gives

$$P(R_1, R_2, \varphi_1, \varphi_2) \simeq \{R_1R_2/[\pi^2(1-d^2)^{1/2}]\} \times \exp\{-(1-d^2)^{-1}[R_1^2+R_2^2 - 2dR_1R_2\cos(\varphi_1+\varphi_2-\delta)]\}. \quad (A5)$$

References

BRAUN, P. B., HORNSTRA, J. & LEENHOUTS, J. I. (1969). *Philips Res. Rep.* **24**, 85-118.

BRUINS SLOT, H. J. & BEURSKENS, P. T. (1984). *Acta Cryst.* A**40**, 701-703.

COCHRAN, W. (1955). *Acta Cryst.* **8**, 473-478.

CROWTHER, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. ROSSMANN, pp. 174-177. New York: Gordon and Breach.

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

DOESBURG, H. M. & BEURSKENS, P. T. (1983). *Acta Cryst.* A**39**, 368-376.

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

HIRSHFELD, F. L. (1968). *Acta Cryst.* A**24**, 301-311.

HUBER, R. (1965). *Acta Cryst.* **19**, 353-356.

GIACOVAZZO, C. (1974). *Acta Cryst.* A**30**, 631-634.

GIACOVAZZO, C. (1983). *Acta Cryst.* A**39**, 685-692.

KARLE, J. (1968). *Acta Cryst.* B**24**, 182-186.

KARLE, J. (1972). *Acta Cryst.* B**28**, 820-824.

LANGS, D. A. (1975). *Acta Cryst.* A**31**, 543-550.

MAIN, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 97-105. Copenhagen: Munksgaard.

NORDMAN, C. E. & NAKATSU, K. (1963). *J. Am. Chem. Soc.* **85**, 353-354.

NORDMAN, C. E. & SCHILLING, J. W. (1970). *Crystallographic Computing*, edited by F. R. AHMED, S. R. HALL & C. P. HUBER, pp. 110-114. Copenhagen: Munksgaard.

SIM, G. A. (1959). *Acta Cryst.* **12**, 813-815.

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

Acta Cryst. (1988). A**44**, 300-308

Application of the Molecular Replacement Method to Multidomain Proteins. 2. Comparison of Various Methods for Positioning an Oriented Fragment in the Unit Cell

BY M. CYGLER* AND W. F. ANDERSON

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

(*Received* 7 *May* 1987; *accepted* 7 *December* 1987)

Abstract

The capabilities of several different methods to determine the correct translation of a model for the application of the molecular replacement method of structure determination to multidomain proteins have been analyzed. The structure of the Fab fragment of the autoimmune anti-poly(dT)-specific antibody HED10 was determined using molecular replacement and provides an example for comparing different methods of determining the correct translation of the model and for evaluating the importance of the parameters used. Expansion to space group $P1$ and phasing with a correctly oriented randomly positioned model was found to be superior to either the Crowther-Blow translation function [Crowther & Blow (1967). *Acta Cryst.* **23**, 544-548] or a brute-force search when only a small part of the molecule was used as a model.

Introduction

In a previous paper (Cygler & Anderson, 1988) we have discussed the application of the rotation function (RF) to determine the orientation of fragments of a multidomain protein in a crystal. As an example of such a protein we have used the anti-poly(dT) immunoglobulin Fab fragment HED10 (Cygler, Boodhoo, Lee & Anderson, 1987). The Fab fragment is composed of two relatively rigid domains, the variable (V) and constant (C) domains, connected by two short polypeptide links that provide flexibility of the domain arrangement in the intact molecule (Amzel & Poljak, 1979; Table 1 of Cygler & Anderson, 1988). In the case of the HED10 Fab fragment the RF solutions have been obtained for V and C domains separately. Here we would like to present the various methods that we have applied to determine the position of these fragments in the unit cell, compare the effectiveness of these approaches and investigate what is the minimum size of the fragment that can be successfully used for such a purpose.

* Present address: Biotechnology Research Institute, National Research Council of Canada, Montreal, Quebec H4P 2R2, Canada.

0108-7673/88/030300-09$03.00

Several different functions with the property of reaching an extremum for the correct position within the unit cell of the properly oriented model have been proposed for the solution of the translation problem. They have recently been reviewed and compared by Beurskens, Gould, Bruins Slot & Bosman (1987). In earlier approaches the position of the molecule was independently determined relative to each of the space-group symmetry elements either in reciprocal space (*e.g.* Vand & Pepinsky, 1956; Tollin, 1966; Crowther & Blow, 1967; Karle, 1972; Langs, 1975) or in direct space (*e.g.* Huber, 1965; Braun, Hornstra & Leenhout, 1969; Nordman & Schilling, 1970) with the effect of reducing the dimensionality of the problem from three to one or two dimensions. A more direct but computationally more demanding approach is a search for an extremum of an agreement index or a correlation coefficient between observed structure amplitudes and those calculated from the model (Cutfield, Cutfield, Dodson, Dodson & Sabesan, 1974; Nixon & North, 1976; Rae, 1977; Harada, Lifchitz, Berthou & Jolles, 1981; Rius & Miravitlles, 1986; Fujinaga & Read, 1987). In recent years the latter methods have become more easily accessible to crystallographers owing to the greatly increased power of modern computers.

Another approach that we describe here is very similar to standard Fourier methods for completing a structure starting from a portion of the structure (*e.g.* Stout & Jensen, 1968). In the present case, the problem is to determine the position of the molecule relative to the symmetry elements and the origin of the proper space group. Consequently, the crystal symmetry is reduced to $P1$ and an electron density map calculated with phases derived from one correctly oriented but arbitrarily positioned molecule. This electron density map is then used for a direct-space search for the symmetry-related molecules. In favorable circumstances symmetry-related molecules can be recognized and the position of symmetry elements determined. We show here that this simple approach can be successfully extended to proteins with the use of as little as 12% of the total unit-cell contents for phasing. Expansion to the triclinic system, coupled with the requirement that about 12% of the unit-cell contents be used for phasing, limits the application of this method to space groups with no more than eight asymmetric units. Knowledge of the orientation of the protein molecule or its fragment can also be utilized in this manner to determine heavy-atom sites in isomorphous derivatives.

The application of the various methods to the positioning of oriented molecular fragments was evaluated with the use of crystal data for the HED10 immunoglobulin Fab fragment (Cygler, Boodhoo, Lee & Anderson, 1987).

Results

Since the best results in determining the orientation of HED10 Fab domains were obtained with the McPC603 Fab fragment (Segal, Padlan, Cohen, Rudikoff, Potter & Davies, 1974; Satow, Cohen, Padlan & Davies, 1986) as a model (Cygler & Anderson, 1988), this Fab was also used for the purpose of positioning the molecule within the unit cell. The crystals of the HED10 Fab fragment belong to the space group $P2_1$; thus only positioning along x and z axes is required. When, however, a relative position of two fragments of the molecule is considered, the difference in their y coordinates also has to be determined.

A number of independent methods were used to determine the positions of the V and C domains of the Fab molecule within the unit cell. They were: (1) correlation-coefficient and R-factor searches; (2) the Crowther-Blow translation function; and (3) expansion to space group $P1$ and phasing based on the correctly oriented fragment in a random position.

Maximizing the correlation coefficient

Various methods of solving the translation problem have been proposed in the literature (*e.g.* Tollin, 1966; Crowther & Blow, 1967; Nixon & North, 1976; Harada, Lifchitz, Berthou & Jolles, 1981; Langs, 1975; Langs, 1985). The most straightforward, although computationally intensive, is the comparison of observed structure factors with those calculated from the model (Cutfield *et al.*, 1974; Nixon & North, 1976; Rae, 1977; Fujinaga & Read, 1987). Two indicators are generally used to assess the agreement, the standard crystallographic R factor and the correlation coefficient, defined as

$$\text{corr} = \frac{\sum (x - \langle x \rangle)(y - \langle y \rangle)}{[\sum (x - \langle x \rangle)^2]^{1/2} [\sum (y - \langle y \rangle)^2]^{1/2}}$$

where $x = F_o$ and $y = F_c$ or $x = F_o^2$ and $y = F_c^2$.

For the calculation of an R factor one requires that the observed and calculated structure factors are on the same scale. The derived value is very sensitive to errors in scaling. The estimate of the scale factor obtained from the Wilson-type statistics, based on data of limited resolution, is not very accurate. The usual method of estimating K for the R-factor calculations is to divide the data into a few resolution ranges and choose the scale for each shell as $K = \sum F_c / \sum F_0$ (Cutfield *et al.*, 1974). This is an adequate method when the model contains a similar number of atoms to the unknown protein molecule; however it underestimates the value of K when only a fragment of the protein is used as a model.

Another potential drawback of using the R factor lies in the fact that for a poor model one expects from

Table 1. *Results of translation searches (6–4 Å resolution) with BRUTE*

Domain	Step	$\Delta\theta$* (°)	Δt* (Å)	Type of search	Maximum correlation coefficient	Peak/r.m.s.
V	1	3·0	1·0	Global x, z local rot	0·241	4·3
V	2	1·0	1·0	Local x, z, rot	0·258	—
C	3	3·0	1·0	Global x, z local rot	0·185	2·9
C	4	1·0	1·0	Local x, z, rot	0·194	—
V, C	5	0·0	0·5	V stationary C global y	0·458	—
V_L, C_L, V_H, C_H1†	6	0·75	0·25	Local x, y, z, rot	0·531	—

Final parameters‡

Domain	α (°)	β (°)	γ (°)	Δx (Å)	Δy (Å)	Δz (Å)
V_L	114·0	102·0	260·2	26·5	0·0	15·5
V_H	108·1	102·5	256·2	26·25	0·25	17·0
C_L	84·4	99·5	257·5	29·25	0·6	17·1
C_H1	79·5	102·0	257·3	29·25	1·0	16·0

* $\Delta\theta$ is the step size of additional rotations along the x, y and z axes applied to the model. Combinations of three rotations ($-\Delta\theta$, 0, $\Delta\theta$) along each axis and their superpositions were included in local searches (27 grid points). Δt is the translational step size that was used. Local searches span ±4 steps in each direction (729 grid points).

† Each unit in turn underwent the local x, y, z and rotation searches while the others remained at the previously determined positions. The process was repeated to be sure no further changes in parameters occurred.

‡ α, β, γ are Euler rotation angles according to Crowther's (1972) notation, Δx, Δy and Δz are the components of the translation vector.

practical experience an R factor in the range of 50% for the correct position of the model, not very far from the value of 58·6% for a random distribution of atoms. The expected variation of the R factor as a function of model position is thus rather small.

The correlation coefficient, on the other hand, is free from errors in scaling. Also, the expected correlation value is in the range of 0·2–0·4, far from the value of 0 for a random distribution of atoms.

We have used a search-type calculation ('brute-force approach') to determine the position of the model within the unit cell. The model, oriented according to the rotation matrix determined previously, is moved in small steps within the area of the unit cell that corresponds to one permissible origin. For each position one calculates the crystallographic R factor and/or the correlation coefficient between F_o^2 and F_c^2. The computer program *BRUTE* (Fujinaga, 1986) was used for these calculations. It is well known that the peak in the translation function is sensitive to small misorientations of the model and that the poorer the starting model the more important is the correct orientation (Dodson, 1985). Consequently the above translation function was calculated repeatedly for a series of slightly different orientations of the model. Each new orientation was obtained by applying an additional rotation, $\Delta\theta$, of $\pm 3°$ about the x, y or z axes, or any combination of such rotations (Read, 1986). This procedure of a six-dimensional search was similar to that employed by Dijkstra, van Nes, Kalk, Brandenburg, Hol & Drenth (1982), with one important exception. Instead of using it after the translation was determined, merely to find a better starting position of the molecule for the subsequent refinement, it was used concomitantly with the translation search. The advantage, at the cost of longer computations, lies in the possibility of correcting for small misorientations while the translation is being determined, and in this way improving the chances for the success of the translation search. In at least one case (*Streptomyces griseus* trypsin; Read, 1986) this procedure was essential in determining the correct position of the molecule.

Data in the 6–4 Å resolution shell (2581 reflections) were used to calculate the correlation coefficient as a function of the shift in x and z for the V domain and the C domain (Table 1). The highest peak for the V domain (corr = 0·24) was over 4σ above the mean value of the map, while for the C domain (corr = 0·19) it was 3σ above the mean. As would be expected, since the V- and C-domain coordinates were taken from the intact McPC603 coordinate set, the x and z translations for V (26·5, 16·0 Å) and C (29·5, 17·0 Å) domains were quite similar. By iterative local six-dimensional searches, with decreasing angular and translational step size, parameters were found for each domain that maximized the correlation coefficient. The only undetermined parameter at this point was the relative translation between the domains along the y axis. The position of the origin along the y axis was fixed by equating the y component of the V-domain translation to zero. The correlation coefficient for the model including both the V and C domains without the connecting regions of the polypeptide chains was calculated as a function of the translation of the C domain along the y axis, yielding the maximum correlation coefficient of 0·46 for 0·5 Å relative translation. The occurrence of single peaks in the translation functions for the C and V

Table 2. *Results of translation searches (BRUTE) with single domains of Fab McPC603 as models*

The meaning of the symbols is the same as in Table 1.

Domain	Resolution (Å)	α (°)	β (°)	γ (°)	Δx (Å)	Δz (Å)	Highest correlation coefficient	Correct solution	Second-highest correlation coefficient	Correct solution
V_L	6-4	114·0	102·0	260·2	25	20	0·140	No	0·135	Yes
C_L	6-4	84·4	99·5	257·5	29	17	0·159	Yes	0·135	No
V_H	6-4	108·1	102·5	256·2	8	12	0·142	No	0·135	Yes
V_H	8-4	108·1	101·8	256·0	27	17	0·143	Yes	0·135	No
V_H	10-4	108·1	101·8	256·0	7	16	0·143	No	0·135	Yes
C_H1	6-4	79·5	102·0	257·3	20	15	0·100	No	0·095	No

domains corresponding to a very similar translation vector left no doubt as to the correctness of the solution.

At this stage each of the four units of the Fab molecule was treated as an independent rigid body. Each was in turn slightly rotated (see Table 1) and translated (within a few ångströms) to maximize the correlation coefficient while the others were fixed. The procedure was reiterated with decreasing rotational and translational step size and led to an increase of the correlation coefficient to 0·53. Final rotation and translation parameters are shown in Table 1.

It is of interest to determine if a smaller fragment can also be properly positioned with this brute-force approach. We have repeated the translation search calculations for each one of the individual V_L, V_H, C_L and C_H1 domains using the final rotation parameters (Table 1). The results are summarized in Table 2. Of these four domains only the C_L domain produced an easily detectable solution with the correlation coefficient being 0·025 higher than the next-highest value. This result is consistent with the fact that the HED10 and McPC603 C_L domain would be expected to have the greatest amino-acid sequence homology. For the V_H domain the results depended on the resolution shell of the data used in the calculations. Only with the 8–4 Å shell did the correct solution have the highest correlation coefficient. When either the 6–4 Å or the 10–4 Å shell were used the highest correlation coefficient did not correspond to the correct translation. For the V_L and C_H1 domains the correct solutions never had the highest correlation coefficient with any of the parameters that were used in the calculations (data not shown).

Crowther–Blow translation function

To compare the results of the brute-force approach with the Crowther–Blow (CB) translation function we have calculated the CB translation function for two models: the C domain of Fab McPC603 which is approximately half of the Fab molecule, and the C_L domain of Fab McPC603 which is a quarter of the Fab molecule. Because there is only one gene coding for the C_L domain of the κ-type light chain in mouse, the sequences of this domain in Fab McPC603 and HED10 should be almost identical. Therefore, use of the C_L domain provides a very favorable example for a model that comprises only one fourth of the unknown molecule.

The model of the C domain was set in a unit cell with *P*1 symmetry and dimensions $a = b = c = 115$ Å, $\alpha = \beta = \gamma = 90°$. Structure factors were computed to a resolution of 4·8 Å. The CB translation function (*MERLOT* package of programs; Fitzgerald, 1988) was calculated with different numbers of reflections included and with different resolution ranges. In some calculations the option of subtracting the contribution from self-vectors (vectors between the atoms of the model) was used. The results for the C domain are shown in Table 3.

The first conclusion from our calculations is that the best discrimination of the correct solution is obtained by using data of intermediate resolution. In the shell 8–4·8 Å the height of the first spurious peak is 72·3% of that for the correct solution, while in the low-resolution shell, 10–6 Å, it is 95·7%.

The second conclusion concerns the number of reflections included in the calculations. Weak reflections are very important for the discrimination of the correct solution. The best results were obtained when all reflections were included in the calculations. It was especially important at low resolution, when a difference of a few percent of the weakest reflections reversed the relative height of the correct and the first spurious peak (rows 1 and 2, Table 3). It also seems that when the orientation of the model is uncertain within a few degrees (mimicked by offsetting the rotational parameters), not removing the self-vectors provides better discrimination between the correct solution and spurious peaks.

Similar calculations were also performed for a model composed of the C_L domain of Fab McPC603. The model was set in a cell with 110 Å edges and structure factors were calculated to a resolution of 4·3 Å. In this case the CB translation function also contained a peak corresponding to the correct solution but it was never the highest feature of the map. In the best case it was number 9 and its height was

Table 3. *Results of Crowther-Blow translation function for Fab McPC603 domains*

The meaning of the symbols is the same as in Table 1.

Resolution	Threshold on F_o	Number of reflections	Rotation of the model α(°)	β(°)	γ(°)	Self-vectors subtracted	Δx (Å)	Δz (Å)	Correct* peak	Next† %
					C domain‡					
Low resolution										
10-6·0	$>3\sigma$	1783	85	100	255	Yes	28·9	17·0	100	95·7
10-6·0	$>5\sigma$	1728	85	100	255	Yes	28·9	17·1	97·5	100
10-6·0	>5000	754	82·9§	100§	256·5§	Yes	—	—	<60	100
Intermediate resolution										
10-4·8	$>2\sigma$	3993	85	100	255	No	28·9	17·3	100	74·0
10-4·8	$>2\sigma$	3993	85	100	255	Yes	29·0	17·0	100	79·3
10-4·8	>1000	3414	85	100	255	Yes	29·1	17·0	100	90·7
10-4·8	>2000	2898	85	100	255	Yes	29·1	17·2	100	99·5
10-4·8	>3000	2419	85	100	255	Yes	29·2	16·9	98	100
10-4·8	>5000	1789	85	100	255	Yes	—	—	<60	100
8-4·8	$>3\sigma$	3465	85	100	255	Yes	29·1	16·8	100	72·3
6-4·8	$>3\sigma$	2161	85	100	255	Yes	28·9	16·9	94	100
Effects of misorientation										
10-4·8	$>2\sigma$	3993	81	100	255	No	29·8	16·8	100	94·7
10-4·8	$>2\sigma$	3993	78	100	255	No	30·5	16·2	95·2	100
10-4·8	$>2\sigma$	3993	81	102	255	No	29·6	16·9	100	86·2
10-4·8	$>2\sigma$	3993	81	102	258	No	29	16·8	100	89·5
10-4·8	$>2\sigma$	3993	87	104	259	No	31·9	17·3	100	70·0
					C_L domain					
10-4·3	$>2\sigma$	5760	84·4	99·5	257·5	Yes	29·0	17·3	69·1	100

* Correct peak (%) is the peak height at the correct translation relative to the highest peak in the map.
† Next % is the peak height for the highest incorrect translation expressed as a percentage of the highest peak in the map.
‡ α, β, γ correspond to the position of the maximum of the rotation function. No interpolation was attempted.
§ This example is shown with the final orientation from *BRUTE* searches, as the most favorable case.

69% of the highest peak in the map. Whether the self-vectors were subtracted or not had little influence on the results. The result of a representative calculation is shown in Table 3.

Expansion to space group P1

A method that is often used successfully in small-molecule crystallography when a partial structure is correctly oriented but misplaced within the unit cell is to expand the data to the space group $P1$ and to calculate an electron density (e.d.) map with phases derived from the partial structure (*e.g.* Karle & Karle, 1971). Such a map reveals, naturally, peaks corresponding to the atoms in the model (Ramachandran & Srinivasan, 1970). Very often it is also possible to distinguish fragments of symmetry-related molecules among the smaller peaks. This is because the positions of maxima in the e.d. map (but not their heights) are relatively insensitive to large errors in the phases of the structure amplitudes (Ramachandran & Srinivasan, 1970). A simple direct-space search for symmetry-related molecules in this e.d. map can be used to determine their positions. The positions of symmetry elements are then derived from the positions of the symmetry-related molecules. This is equivalent to defining the origin. The method often works even when as little as 10–15% of the total unit-cell contents are used. We have tested the usefulness of this method for macromolecules applying two different approaches. The first used the amplitude differences due to the introduction of a heavy atom and the second was based on the native protein amplitudes.

Heavy-atom difference map. We have applied this method to the amplitude differences between a heavy-atom derivative and the native protein. A K_2HgI_4 derivative was particularly suitable because it has one major site. For the same reason, the phases calculated from this derivative are nearly centrosymmetric (apparent space group $P2_1/m$) and not particularly useful in estimating native protein phases. The difference coefficients were expanded to a full hemisphere and the phase for each reflection was calculated from either the V or the C domain, correctly oriented but arbitrarily positioned in the unit cell, assuming $P1$ symmetry. These structure factors were used to calculate difference e.d. maps. In each case there were two peaks above the background. In the difference e.d. map calculated using phases derived from the C domain the relative heights of the two highest peaks were 1·0 and 0·71 (0·44 for the highest background peak) and the fractional y coordinates differed by 0·5 (Table 4). Similar results were obtained with the V domain. The positions of these two peaks

Table 4. *Two symmetry-related positions of the major heavy-atom site of a K_2HgI_4 derivative obtained from a difference map calculated in space group P1 (6Å resolution) phased by an arbitrarily positioned V or C domain of Fab McPC603 that was oriented according to the rotation function peak*

The translation vector to place the model correctly in space group $P2_1$ that is derived from these pairs of sites can be compared with the values in Table 1.

	Site 1			Site 2			Translation		
Domain	x	y	z	x	y	z	Δx (Å)	Δy (Å)	Δz (Å)
V	−0·28	0·00	−0·25	−0·54	0·50	−0·52	26·3	0·0	16·3
C	−0·32	−0·01	−0·27	−0·58	0·46	−0·52	28·9	1·3	16·8

defined the twofold screw axis. Its x and z coordinates provide the translation vector required to position the model relative to an origin on the 2_1 axis. It was also possible to determine the relative translations of the V and C domains along x and z as well as the y axis from the difference in the coordinates of the heavy-atom sites in the difference e.d. maps phased by each domain. The translation vectors (Table 4) obtained in this way were very close to the translation vectors for the two domains obtained by the correlation-coefficient search procedure (Table 1).

Fourier map. In this case the structure-factor amplitudes observed for the native protein were expanded to the full hemisphere and a Fourier map at 3·5 Å resolution was calculated with phases derived from an arbitrarily placed but correctly oriented model of either the V or the C domain as described above. This map had high electron density within the volume of the 'phasing' molecule, while the density was much lower elsewhere. The position of a symmetry-related molecule within this map was determined by a direct-space search. The 'test' molecule used in the search was created by applying the transformation $-x, \frac{1}{2}+y, -z$ to the 'phasing' molecule. This 'test' molecule was then moved within the e.d. map and a score function calculated at each position. We have chosen a very simple score function,

$$S(p, q, r) = \sum_{i=1}^{N} \rho(x_i + p, y_i + q, z_i + r),$$

where the summation is over all N atoms of the test fragment, ρ is the electron density (on an arbitrary scale), (x_i, y_i, z_i) is the e.d. map grid point closest to the position of the ith atom, and (p, q, r) is the translation vector applied to the 'test' molecule. A distance between grid points of 1·2 Å was used (one third of the resolution) in the calculation of the e.d. map. The volume occupied by the molecule used to derive phases must be avoided during searches for two reasons. Firstly, since the electron density within this volume is much higher than elsewhere, the score function corresponding to placing the 'test' molecule in that volume could easily be higher than elsewhere, even though the fit would be very poor. Secondly, one does not expect the two symmetry-related molecultes to overlap. The most straightforward method of achieving the desired effect is to modify the map by setting the electron density at grid points adjacent to atoms of the original phasing model to a large negative number. In order to simplify calculations the electron density was calculated in the range: 0·0 to $1·0+\Delta x$ in the x direction, y_{gc} to $1·0+y_{gc}$ in the y direction and 0·0 to $1·0+\Delta z$ in the z direction, where Δx and Δz are the corresponding dimensions of the search molecule and y_{gc} is the y coordinate of the geometric center of the search molecule. This ensures that the 'test' molecule is present as one undivided entity within the map.

For the correct solution one expects the difference $\Delta y = 0·5$ between the y coordinates of the 'phasing' molecule and the position of the 'test' molecule corresponding to the maximum score. The x and z coordinates of the screw axis relating the 'phasing' molecule and the 'test' molecule can then be calculated. This provides the shift which should be applied to the original molecule to position the 2_1 screw axis at an origin. The results of this procedure are given in Table 5 for the various parts of the Fab McPC603 molecule that were used to generate the starting phases.

Both the V and the C domains oriented according to the final angles from the brute-force approach (Table 1) gave a peak at the expected position that was approximately twice as high as the first spurious peak. This corresponded to a case where only 25% of the unit-cell contents were used to calculate phases. What was found to be very encouraging was the relative insensitivity of this peak to errors in the orientation of the original molecule. Imposing an error of as much as 6·5° in the orientation of the C domain did not affect the rank of the correct solution although the difference between the correct and the first spurious peak decreased (Table 5). The tests with the V domain showed that the result was more sensitive to misorientation.

The calculations were repeated for each of the single domains V_L, V_H, C_L or C_H1 of Fab McPC603 (~12·5% of the total unit-cell contents), oriented according to the final angles of Table 1. For three of them (C_L, V_L and V_H) the highest score S corresponded to the correct solution. For the C_H1 domain the

Table 5. *Results of searches (score function S) for a symmetry-related fragment of the molecule in an electron density map calculated in the space group P1 with phases derived from an arbitrarily positioned molecular fragment*

Fragment	Rotation of model molecule α (°)	β (°)	γ (°)	$\Delta\varphi$* (°)	Score S Correct peak	Highest spurious peak
C	82·4	100·0	255·5	0·0	293	172
C	86·0	103·0	258·0	5·0	267	159
C	87·0	104·0	259·0	6·5	203	146
V	110·4	102·2	258·4	0·0	311	164
V	108·0	105·0	261·0	4·5	165	185
V	106·0	106·0	262·0	7·4	136	196
C_L	84·4	99·5	257·5	0·0	209	132
C_L	85·0	101·0	255·0	3·1	146	118
C_L	85·0	105·0	255·0	6·1	87	111
C_H1	79·5	102·0	257·3	0·0	97	110
V_H	108·1	102·5	256·2	0·0	138	125
V_L	114·0	102·0	260·2	0·0	123	116

* Angular deviation from the correct orientation as determined with *BRUTE* (Table 1).

correct translation corresponded to the second-highest peak of the score function. The peak for the C_L domain was 1·6 times the first spurious peak (Table 5). The correct solution still had the highest score when the orientation error of ~3·0° was introduced. For V_L and V_H domains, where the model is not as good as in the case of the C_L domain, the highest score was only ~10% above the spurious maxima and its height was much more sensitive to orientational errors than was the case for the C_L domain.

This procedure can be used in a more general way to find the position of one molecular fragment in a map phased by another fragment. If the position of the 'phasing' fragment within the unit cell is known, the map is calculated in the proper space group, otherwise *P*1 symmetry is assumed. In such a case the modification of the electron density within the volume occupied by the 'phasing' fragment (and symmetry-related fragments) is especially important since the search is performed in close proximity to the 'phasing' fragment. For example, a correctly oriented C domain was translated within a map phased by the oriented V domain (*P*1 symmetry, map limits estimated from the known arrangement of domains in the Fab fragment) and the maximum score *S* corresponded to the correct relative positions of these two domains.

The effect of a partially correct model

When a molecule with internal flexibility is used as a model for rotation function calculations it is possible to misinterpret the results by assigning one set of rotation angles for the whole molecule. What would be the effect of such a partially correct model on the results of the translation search?

To address this question we have carried out the translation search (*BRUTE*; Fujinaga, 1986) using the intact Fab McPC603 molecule oriented according to the rotation function peak corresponding to only the V domain. This resulted in the C domain being about 35° away from the orientation observed for HED10. Following the same procedure as described above, *i.e.* repeating the calculation of the translation function for slightly misoriented models with $\Delta\theta = 3°$, four peaks with correlation coefficients close to 0·19 were observed at slightly different orientations of the molecule. The peak corresponding to the known correct translation had a slightly higher correlation coefficient than the others. Inspection of intermolecular contacts on a graphics system revealed that all four possible solutions led to some steric overlaps with neighboring molecules (in the complementarity-determining regions of the antibody that are likely to differ from the model), and that there was no obvious indication from this information which was the correct one. When this model of the intact Fab, translated correctly for the V domain, was adjusted as described above by a local six-dimensional search with the molecule divided into four rigid bodies the angular search range was not large enough to correct the misoriented portion of the model. The correlation coefficient was only 0·30 as compared to 0·53 for the correct solution. The e.d. map calculated with coefficients $(2F_o - F_c)$ and with phases derived from this model had continuous density within the V domain which was correctly placed but the density was rather disconnected in the incorrectly placed C domain, indicating possible errors in the placement of this part of the molecule.

The partially correct model was also checked against the heavy-atom difference e.d. map. Sets of phases were calculated (in the space group $P2_1$) for each of the four possible translations. In only one case (the correct translation for the V domain) was a single dominant peak observed. Its coordinates agreed with the heavy-atom position determined independently by solving the difference Patterson function. A difference e.d. map calculated with phases derived from the V or the C domain of that model showed that only the V domain reproduced the heavy-

atom site. This indicates that for a derivative with a heavy-atom site of high occupancy even quite large errors in phases lead to an easily interpretable difference map. A similar experience was reported by Evans (1985). At one stage during the MR process, a correctly oriented molecule of phosphofructokinase was misplaced along the y axis by 7 Å and the difference map for a poor derivative had peaks at many of the predicted mercury positions.

Discussion

Three different approaches were applied to solve the translation problem for the Fab HED10 crystal structure. All of them were successful with models composed of either the V or the C domain of the Fab McPC603 molecule ($\sim 50\%$ of all atoms) but differed in the rate of success with smaller fragments. They also show different sensitivity to the parameters used in the computations, such as the resolution of the data and the number of reflections included.

There are two types of score functions that are currently used in the search procedures based on a systematic translation of the model molecule within the unit cell: the crystallographic R factor and the correlation coefficient between observed and calculated structure factors, computed at every position of the model. It is our experience and that of others (Fujinaga & Read, private communication) that when the global minimum value of the R factor corresponds to a different translation from that of the global maximum of the correlation coefficient, the latter is usually the correct one (that is, if the solution can be detected at all). Differences between these indicators are most evident when the rotation angles are only crudely determined. Any difference between the R factor at the correct position of the molecule and at the global minimum, however, is small. The procedure of concomitant small alterations in the orientation of the model during translation searches alleviates this problem to a large extent and can help to identify the solution to the translational problem in those cases that show strong sensitivity to small misorientations of the model. This procedure has been extremely useful in the solution of the translation problem for *Streptomyces griseus* trypsin (Read, 1986). It was also essential in the identification of the correct translation (along with three others) in our test with a partially correct model.

An accurate placement of the molecule is essential for smooth progress of the restrained least-squares refinement since even small orientational errors lead to large positional errors for the outer parts of the molecule. Fine adjustments in positioning of the molecule are usually achieved by a rigid-body refinement (*e.g. CORELS*; Sussman, Holbrook, Church & Kim, 1977) or, when the molecule may be reasonably approximated as one rigid body, by a local six-dimensional R factor or correlation-coefficient search. Following Fujinaga & Read (1987) we have extended the latter approach to the multi-rigid-body situation of the Fab molecule.

The Crowther–Blow translation function performed well when the model comprised a significant part ($\sim 50\%$) of the unknown molecule. With a good starting model the discrimination of the correct solution from spurious peaks was very good when a broad resolution shell was used and all reflections within that shell were included in the calculations. With only low-resolution data (10–6 Å) a strong dependence of the peak height on the presence of weak reflections was observed (Table 3, rows 1 and 2). In all resolution shells elimination of weak reflections led to poorer discrimination. Tests with small misorientations of the model have shown that errors of the order of 5° still revealed the solution clearly. The low sensitivity of the CB function to errors in the orientation depends very much on the quality of the starting model and is rather difficult to generalize. When, however, the orientation of the model was only approximate, not subtracting the self-vectors of the model molecule resulted in better discrimination.

The performance of the CB translation function was rather poor when only a quarter of the unknown molecule was taken as a model. Calculations for the C_L domain of Fab McPC603 showed that the proper translation corresponds, at best, to the ninth peak of the CB function. Since this domain, of the four domains of Fab McPC603, shows the best structural similarity to the corresponding Fab HED10 domain no other tests were attempted.

The method based on expansion to space group $P1$ served two purposes. When applied to the amplitude differences between a heavy-atom derivative and the native protein it allowed the determination of the positions of heavy atoms and the translation of the molecule at the same time. Thus, the results of the rotation function alone can be utilized for the purpose of sorting out the heavy-atom sites of an isomorphous derivative. Application of the phases to the native amplitudes allowed the determination of the positions of symmetry elements and revealed the necessary translation of the model molecule. Modification of the electron density within the volume of the 'phasing' molecule to a large negative number also eliminates possible false solutions that correspond to a significant overlap of symmetry-related molecules. This procedure would then also possess some of the qualities of a packing-function analysis. This approach can easily be extended to the determination of the relative translations of fragments of a molecule.

The ability of each of the methods tested above to arrive at the correct solution is affected by the size of the starting model. With a model corresponding to a quarter of the unknown protein the CB translation function failed to yield the solution while the brute-

force approach and expansion to the space group $P1$ met with partial success. The brute-force approach yielded a clearly recognized solution for the C_L domain and was also able to produce, under certain conditions (8-4 Å resolution), the highest correlation coefficient for the correct translation of the V_H domain. The other two domains, V_L and C_H1, led to a multitude of possible translations with the correct one being somewhat down the list. The latter domain also gave the lowest correlation coefficient of any of the domains. The third method, expansion to the space group $P1$, was the most successful when dealing with small parts of the molecule as a model. In this case the C_L, V_L and V_H domains of Fab McPC603 resulted in the highest score function for the correct translation, with the former domain giving a very clear signal. For the C_H1 domain the correct solution gave the second-highest score. What shows the sensitivity of this approach is the fact that it was possible in the same way to determine the translation of the V_H domain relative to the C_L domain (data not shown).

Each of these methods is differently affected by the number of reflections included in the calculations. The most sensitive seems to be the CB function, for which all reflections in a given resolution range should be included. For the brute-force approach weak reflections (or a random sample of all reflections) should be included for proper statistics. In the Fourier method the weak reflections are far less important as their impact on the e.d. map is not large. Finally, there is the question of the resolution shell of the data to be included in the calculations. In this respect the brute-force approach is superior to the other two methods. The correct answer can be obtained with a narrow shell of medium resolution [*e.g.* 5-4 Å (Read, 1986; Fujinaga & Read, 1987)], while the other two methods require larger shells of data (10-4 or 8-4 Å).

Finally, our tests with a partially correct model have shown (see also Evans, 1985) that the correct determination of heavy-atom sites from a difference map phased by the model is not in its own right an adequate indication that the model does not contain gross errors. Even a model with 50% of the molecule placed incorrectly can identify heavy-atom sites satisfactorily.

We acknowledge the help of D. Bacon with some programs used in this study, P. M. D. Fitzgerald for access to the *MERLOT* package of programs and especially R. Read, M. Fujinaga and Alastair Muir for helpful discussions. The financial assistance of the Medical Research Council of Canada to the Group in Protein Structure and Function is gratefully acknowledged.

References

AMZEL, L. M. & POLJAK, R. J. (1979). *Annu. Rev. Biochem.* **48**, 961-997.

BEURSKENS, P. T., GOULD, R. O., BRUINS SLOT, H. J. & BOSMAN, W. P. (1987). *Z. Kristallogr.* In the press.

BRAUN, P. B., HORNSTRA, J. & LEENHOUT, J. I. (1969). *Philips Res. Rep.* **24**, 85-118.

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

CUTFIELD, J. T., CUTFIELD, S. M., DODSON, E. J., DODSON, G. G. & SABESAN, M. (1974). *J. Mol. Biol.* **87**, 23-30.

CYGLER, M. & ANDERSON, W. F. (1988). *Acta Cryst.* A**44**, 38-45.

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

DIJKSTRA, B. W., VAN NES, G. J. H., KALK, K. H., BRANDENBERG, N. P., HOL, W. G. J. & DRENTH, J. (1982). *Acta Cryst.* B**38**, 793-799.

DODSON, E. (1985). *Molecular Replacement.* Proceedings of the Daresbury Study Weekend, 15-16 February 1985, edited by P. A. MACHIN, pp. 33-45. Warrington: SERC Daresbury Laboratory.

EVANS, P. R. (1985). *Molecular Replacement.* Proceedings of the Daresbury Study Weekend, 15-16 February 1985, edited by P. A. MACHIN, pp. 53-56. Warrington: SERC Daresbury Laboratory.

FITZGERALD, P. M. D. (1988). *J. Appl. Cryst.* **21**. In the press.

FUJINAGA, M. (1986). PhD Thesis. Univ. of Alberta, Canada.

FUJINAGA, M. & READ, R. (1987). *J. Appl. Cryst.* **20**, 517-521.

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

HUBER, R. (1965). *Acta Cryst.* **19**, 353-356.

KARLE, I. L. & KARLE, J. (1971). *Acta Cryst.* B**27**, 1891-1898.

KARLE, J. (1972). *Acta Cryst.* B**28**, 820-824.

LANGS, D. A. (1975). *Acta Cryst.* A**31**, 543-550.

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

NIXON, P. E. & NORTH, A. C. T. (1976). *Acta Cryst.* A**32**, 320-325.

NORDMAN, C. E. & SCHILLING, J. W. (1970). In *Crystallographic Computing*, edited by F. R. AHMED, S. R. HALL & C. P. HUBER, pp. 110-114. Copenhagen: Munksgaard.

RAE, A. D. (1977). *Acta Cryst.* A**33**, 423-425.

RAMACHANDRAN, G. G. & SRINIVASAN, R. (1970). *Fourier Methods in Crystallography.* New York: John Wiley.

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

RIUS, J. & MIRAVITLLES, C. (1986). *Acta Cryst.* A**42**, 402-404.

SATOW, Y., COHEN, G. H., PADLAN, E. A. & DAVIES, D. R. (1986). *J. Mol. Biol.* **190**, 593-604.

SEGAL, D. M., PADLAN, E. A., COHEN, G. H., RUDIKOFF, S., POTTER, M. & DAVIES, D. R. (1974). *Proc. Natl Acad. Sci. USA*, **71**, 4298-4302.

STOUT, G. H. & JENSEN, L. H. (1968). *X-ray Structure Determination. A Practical Guide.* London: Macmillan.

SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S. H. (1977). *Acta Cryst.* A**33**, 800-804.

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

VAND, V. & PEPINSKY, R. (1956). *Z. Kristallogr.* **108**, 1-14.

Acta Cryst. (1988). A44, 309-320

Analysis of Solvent Structure and Hydrogen Exchange in Proteins on the Basis of Neutron Diffraction Data from Deuterated and Hydrogenous Crystals

BY ROBERT W. HARRISON* AND ALEXANDER WLODAWER*

Center for Chemical Physics, National Bureau of Standards, Gaithersburg, MD 20899, USA

AND LENNART SJÖLIN

Department of Inorganic Chemistry, Chalmers Institute of Technology, Göteborg, Sweden

(*Received* 4 *August* 1987; *accepted* 10 *December* 1987)

Abstract

A method has been developed to determine the structure of bound solvent and the positions of exchanged hydrogens in proteins, on the basis of neutron diffraction from hydrogenous and deuterated crystals. In this method phases for the hydrogenous and for the deuterated model are refined simultaneously, and an average model is imposed in the volume occupied by non-hydrogen atoms. The densities in the areas of bulk solvent are replaced by their average values, while no modifications are performed in the vicinity of ordered solvents and potentially exchangeable hydrogens. The method was tested on 1·8 Å neutron diffraction data collected from two crystals of bovine pancreatic trypsin inhibitor, one of them deuterated and the other hydrogenous. Significant improvement was observed for the densities corresponding to many partially occupied solvent sites, as well as to partially exchanged hydrogens. The algorithm presented here has been compared with a different approach published recently by Shpungin & Kossiakoff [*Methods Enzymol.*(1986), **127**, 329-342].

Introduction

Protein crystals contain large numbers of water molecules in fully or partially occupied sites, usually in close proximity to the protein surface. Water positions are always assigned during refinement of protein structures at high resolution, but their number varies considerably in different structural investigations. For example, 40 waters were assigned in a very conservative neutron study of carbon monoxymyoglobin (Hanson & Schoenborn, 1981), while 316 solvents were placed in the course of an X-ray refinement of oxymyoglobin (Phillips, 1980). Detailed comparisons of solvent positions have been performed for the same structures refined independently in different laboratories (Wlodawer, Borkakoti, Moss & Howlin, 1986), for the same proteins in different crystal forms (Wlodawer, Deisenhofer & Huber, 1987; Wlodawer, Nachman, Gilliland, Gallagher & Woodward, 1987), or for two crystallographically independent molecules in the same unit cell (Blevins & Tulinsky, 1985). All of these investigations were in agreement that while the majority of solvent positions present in these models were real, some were probably artefacts. In the absence of confirming information, positions of partially occupied waters, or of waters not forming clear hydrogen bonds with the protein, must be treated with a degree of skepticism.

It has been suggested (Schoenborn, 1969; Wlodawer, 1982; Kossiakoff, 1983) that neutron diffraction should be a method of choice for crystallographic investigations of water structure in macromolecular crystals, especially if D_2O is used to replace H_2O. This is due to the much higher relative contribution of deuterium atoms to the overall scattering, since the deuterium scattering length is almost the same as that of oxygen. In practice, however, our results of neutron refinements of solvent structure have been disappointing. Neutron data are usually of poorer quality and extend to lower resolution than the X-ray data from the same crystals (Savage & Wlodawer, 1986; Wlodawer, Walter, Huber & Sjölin, 1984). This is caused by two unfavorable properties of the neutron diffraction technique. The available fluxes of neutrons are five to six orders of magnitude lower than the fluxes of X-rays, and the incoherent scattering of unexchanged hydrogens makes the intrinsic signal-to-noise ratio much lower in neutron diffraction. The latter problem could be circumvented only by the use of completely deuterated proteins (in which even the non-exchangeable hydrogens have been substituted by deuteriums), but no such studies have been reported so far. While the increase of resolution beyond about 2Å does not necessarily improve solvent maps, high-resolution data help in obtaining better phases and thus lead indirectly to easier interpretation of Fourier maps.

Neutron diffraction has been utilized in a number of hydrogen-exchange studies of protein single

* Current address: Crystallography Laboratory, NCI - Frederick Cancer Research Facility, BRI - Basic Research Program, PO Box B, Frederick, Maryland 21701, USA.

0108-7673/88/030309-12$03.00