Acta Cryst. (1998). A**54**, 617–625

The Translation Problem in Molecular Replacement Techniques. I. About the Role of Triplet Invariants

C. Giacovazzo,[a]* L. Manna,[a] D. Siliqi,[a]† M. Bolognesi[b] and M. Rizzi[b]

[a]*Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and* [b]*Dipartimento di Genetica e Microbiologia, Sezione di Cristallografia, Università, Via Abbiategrasso 207, 27100 LPH, Italy. E-mail: crisg01@area.ba.cnr.it*

(*Received 15 February 1997; accepted 19 February 1998*)

Abstract

The case of a well oriented but randomly positioned molecule has been treated in a pioneering paper by Main (1976) [in *Crystallographic Computing Techniques*, edited by F. R. Ahmed. Copenhagen: Munksgaard]. The formula proved quite effective for small molecules but in its original form is inadequate for solving the translation problem in molecular replacement techniques applied to proteins. The Main formula has been suitably modified: applications to test structures show that the use of direct methods may be a valid alternative to the widely used translation functions.

1. Symbols and abbreviations

$F_{\mathbf{h}}$: structure factor of the protein with vectorial index $\mathbf{h}$
ϕ_h: phase of $F_{\mathbf{h}}$
$f_j(\mathbf{h})$: scattering factor of the jth atom
$\mathbf{C}_s \equiv (\mathbf{R}_s, \mathbf{T}_s)$: sth symmetry operator. $\mathbf{R}_s$ is the rotational part, $\mathbf{T}_s$ the translational part.
m: order of the point group of the space group
N: number of atoms in the unit cell
N_f: number of molecular fragments (symmetry independent) with unknown position and fixed orientation.
n_i: number of atoms in the ith molecular fragment
q: number of atoms (symmetry equivalents included) whose positions are completely unknown
$\sum_q(\mathbf{h}) = \sum_{j=1}^{q} f_j^2(\mathbf{h})$: scattering power of the q atoms with completely unknown position

$$\sum\nolimits_N(\mathbf{h}) = \sum_{j=1}^{N} f_j^2(\mathbf{h})$$

$$\sum\nolimits_{3q}(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) = \sum_{j=1}^{q} f_j(\mathbf{h}_1) f_j(\mathbf{h}_2) f_j(\mathbf{h}_3)$$

$$\sum\nolimits_{3N}(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) = \sum_{j=1}^{N} f_j(\mathbf{h}_1) f_j(\mathbf{h}_2) f_j(\mathbf{h}_3)$$

$$\sigma_i = \sum_{j=1}^{N} Z_j^i,$$

where Z_j is the atomic number of the jth atom
ε_h: Wilson's factor responsible for the enhancement or depression of the intensity of certain subsets of reflections due to particular symmetry elements
$\Phi = \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3}$ with $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$
$I_i(x)$: modified Bessel function of order i
$D_i(x) = I_i(x)/I_0(x)$.

2. Introduction

Cochran (1955) estimates of triplet invariant phases are not sufficiently accurate for macromolecular crystallography. It was argued that the estimates would greatly improve if some means is found for making use of available prior information. The knowledge of the correct orientation of one or more groups of atoms randomly positioned was exploited by Main (1976) and reconsidered by Giacovazzo *et al.* (1988), who obtained additional probabilistic formulas for polar space groups. The method has never been systematically applied (at least to the knowledge of the authors) to macromolecules: a recent contribution by Langs *et al.* (1995) for a phase-invariant translation function aims at determining the heavy-atom position for single isomorphous replacement data.

The first aim of this paper is to show that a direct-methods procedure can be designed that preserves its efficiency even when applied to proteins. It might be worthwhile mentioning that the problem of locating a well oriented molecule is of primary importance in molecular replacement techniques (see Rossmann, 1990, and literature quoted therein). While rotation functions frequently succeed in finding the correct orientation of the molecule, the translation functions may show many maxima and the correct translation may not correspond to the largest one. A thorough review of the literature by Beurskens *et al.* (1987) is highly recommended. We want

† Permanent address: Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana University, Tirana, Albania.

to show that a direct procedure based on triplet estimates can be competitive with the most widely used translation functions.

3. The Main formula

Let us divide the crystal structure into two parts: the first includes N_f molecular fragments with known orientation and their symmetry equivalents. The second comprises the q atoms (symmetry equivalents included) whose positions are completely unknown. The conditional probability distribution function of the triplet phase Φ was stated in a pioneering paper by Main (1976):

$$P(\Phi) \approx [2\pi I_0(Q)]^{-1} \exp[Q\cos(\Phi - \Theta)],$$

where

$$Q = 2|E_{M\mathbf{h}_1} E_{M\mathbf{h}_2} E_{M\mathbf{h}_3}| \left[\frac{Q'^2 + Q''^2}{\langle|F_{\mathbf{h}_1}|^2\rangle_M \langle|F_{\mathbf{h}_2}|^2\rangle_M \langle|F_{\mathbf{h}_3}|^2\rangle_M}\right]^{1/2}, \tag{1}$$

$$Q' = \Re\left\{\sum_{i=1}^{N_f}\sum_{s=1}^{m} g_{is}(\mathbf{h}_1) g_{is}(\mathbf{h}_2) g_{is}(\mathbf{h}_3) + \sum\nolimits_{3q}(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)\right\},$$

$$Q'' = \Im\left\{\sum_{i=1}^{N_f}\sum_{s=1}^{m} g_{is}(\mathbf{h}_1) g_{is}(\mathbf{h}_2) g_{is}(\mathbf{h}_3)\right\},$$

$$g_{is}(\mathbf{h}) = \sum_{j=1}^{n_i} f_j(\mathbf{h}) \exp(2\pi i \mathbf{h}\mathbf{C}_s\mathbf{u}_j),$$

$$\tan\Theta = Q''/Q',$$

$\Re\{\ldots\}$ and $\Im\{\ldots\}$ stand for the real and imaginary parts of $\{\ldots\}$, respectively, $\mathbf{u}_j$ is the trial position of the jth atom (belonging to the oriented fragment) and

$$\langle|F_{\mathbf{h}}|^2\rangle_M = \varepsilon_{\mathbf{h}}\left\{\sum_{i=1}^{N_f}\sum_{s=1}^{m} |g_{is}(\mathbf{h})|^2 + \sum\nolimits_q(\mathbf{h})\right\} \tag{2}$$

is the expected (on the basis of the prior information) value of $|F_{\mathbf{h}}|^2$. $E_{M\mathbf{h}} = F_{\mathbf{h}}/\langle|F_{\mathbf{h}}|^2\rangle_M^{1/2}$ is the normalized structure factor.

4. A tentative procedure

In macromolecular crystallography, the wide use of molecular replacement methods suggests the following scenario: a model molecule, similar to that under study, is oriented by some rotation function. Then the correct translation is searched for; once this has been found, a refinement process starts to modify the electron density of the translated model molecule into the electron density of the molecule under study. In this case, the translation problem has peculiar features: we are dealing with a model molecule that may be weakly correlated with the molecule under study and its orientation may be roughly accomplished. A direct phasing procedure designed for this type of problem should:

(*a*) deal with a large number of reflections, and therefore a large number of triplets;

(*b*) be efficient in the calculation of the terms Q' and Q'' for numerous triplets;

(*c*) provide the correct solution to the translation problem in a reasonable time.

The steps of our trial procedure may be described as follows:

(i) The observed structure factors $|F|$ of the crystal structure under study are scaled by a Wilson plot. Let K and B be the estimated scaling and overall isotropic thermal factors.

(ii) The structure factors $F_{M\mathbf{h}}$ corresponding to the correctly oriented (but wrongly positioned) model molecule are calculated:

$$F_{Mh} = \sum_{i=1}^{n_i}\sum_{s=1}^{m} g_{is}(\mathbf{h}). \tag{3}$$

In (3), we use the B factor obtained in step (i).

(iii) The scaled $|F|$ are normalized:

$$E_{M\mathbf{h}} = F_{\mathbf{h}}/\langle|F_{\mathbf{h}}|^2\rangle_M^{1/2}.$$

Because of reasons which will be described later, two types of normalized structure factors will be calculated, called E_M when $\langle|F_{\mathbf{h}}|^2\rangle$ is fixed by equation (2) and E_W when $\langle|F_{\mathbf{h}}|^2\rangle = \varepsilon\sum_N$. In the latter case, the prior information on the model molecule is overlooked.

(iv) Reflections are arranged in decreasing order of $|E|$. A relatively small number of reflections (say the NLAR with the largest $|E|$ values) are selected, among which triplet relations are found.

(v) A random approach is chosen (Baggio *et al.*, 1978) to which the weighted tangent formula

$$\tan(\phi_{\mathbf{h}}) = \frac{\sum w_{\mathbf{k},\mathbf{h}-\mathbf{k}} Q_{\mathbf{k},\mathbf{h}-\mathbf{k}} \sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}} + \Theta_{\mathbf{k},\mathbf{h}-\mathbf{k}})}{\sum w_{\mathbf{k},\mathbf{h}-\mathbf{k}} Q_{\mathbf{k},\mathbf{h}-\mathbf{k}} \cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}} + \Theta_{\mathbf{k},\mathbf{h}-\mathbf{k})}} = \frac{A_{\mathbf{h}}}{B_{\mathbf{h}}} \tag{4}$$

is applied. The reliability of the phase estimate is

$$\alpha_{\mathbf{h}} = (A_{\mathbf{h}}^2 + B_{\mathbf{h}}^2)^{1/2}. \tag{5}$$

The weighting scheme is designed to drive phases towards values that minimize the difference between α and $\langle\alpha\rangle$ (Hull & Irwin, 1978; Altomare *et al.*, 1994).

(vi) The correct solution is chosen among the various trials by suitable figures of merit (see §8) and is used as a seed for phasing the remaining reflections. Batches of about 200 reflections, chosen in decreasing order of $|E|$, are progressively phased *via* a phase-extension procedure from the NLAR reflections.

The above procedure is similar to that recently described in a series of papers integrating direct methods with isomorphous replacement techniques (see Giacovazzo *et al.*, 1996, and literature quoted therein).

The sequence of the various steps agrees with a general strategy: subdivide the set of reflections to phase into small batches in order to avoid the simultaneous calculations of several tens of millions of triplets, their cumbersome management by the tangent formula, and the need for large storage and computing times.

A further point that deserves to be noticed concerns the calculation of the Q term in (1). Since triplets involve both standard and symmetry-equivalent reflections, we have to calculate functions like $g_{is_1}(\mathbf{hR}_{s_2})$ on varying indices $\mathbf{h}$, s_1 and s_2. The task may not be trivial owing to the fact that: (*a*) for each triple $\mathbf{h}, s_1, s_2$ the contribution of a large number of atoms has to be calculated; (*b*) the same calculation may be repeated several times (each $\mathbf{h}$ may enter in hundreds and sometimes in thousands of triplets). The problem may be simplified if one observes that

$$g_{is_1}(\mathbf{hR}_{s_2}) = \sum_{j=1}^{n_i} f_j(\mathbf{h}) \exp 2\pi i \mathbf{h}(\mathbf{R}_{s_2}\mathbf{R}_{s_1}\mathbf{u}_j + \mathbf{R}_{s_2}\mathbf{T}_{s_1}). \quad (6)$$

Defining the operator $\mathbf{C}_{s_3} \equiv (\mathbf{C}_{s_2}\mathbf{C}_{s_1}) \equiv (\mathbf{R}_{s_2}\mathbf{R}_{s_1}, \mathbf{R}_{s_2}\mathbf{T}_{s_1} + \mathbf{T}_{s_2})$, (6) may be written as

$$g_{is_1}(\mathbf{hR}_{s_2}) = \exp(-2\pi i \mathbf{hT}_{s_2}) g_{is_3}(\mathbf{h}). \quad (7)$$

According to (7), we only need to calculate $g_{is}(\mathbf{h})$, $s = 1, 2, \ldots, m$. In our procedure, they are stored for the various $\mathbf{h}$ involved in the triplet calculations: when $g_{is_1}(\mathbf{hR}_{s_2})$ is needed, its value may be obtained *via* the multiplication table of the symmetry-operator group. The last remark concerns the role of the terms $\sum_{3q}(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)$ and $\sum_q(\mathbf{h})$ in (1) for macromolecular crystallography. $\sum_{3q}$ represents the Cochran contribution to the reliability factor of that part of the asymmetric unit which is completely unknown: if it was non-negligible, it would drive the triplet phases towards 2π. Since the model molecule usually represents a large part of the protein molecule, the value of $\sum_{3q}$ is frequently negligible with respect to $\sum_{i=1}^{N_f}\sum_{s=1}^{m} g_{is}(\mathbf{h}_1)g_{is}(\mathbf{h}_2)g_{is}(\mathbf{h}_3)$. Therefore, it (as well as $\sum_q$) may be neglected in all our calculations: accordingly, estimated triplet phases will be almost uniformly distributed between $-\pi$ and π.

5. The test structures

We will apply our procedure to four test structures: code name, space group and other useful data are in Table 1. Molecular replacement techniques have been an essential tool for their solution: a model structure was rotated into the correct position by application of some rotation functions, then some translation function was used to correctly locate it. In our tests, the same model molecule will be used for solving the rotational problem, but a translation search will be performed by direct methods. It is very useful to briefly characterize the search model with respect to the test structure.

M-FABP (recombinant human-muscle fatty-acid-binding protein): $P2_12_12_1$, $a = 35.400$, $b = 56.700$, $c = 72.700$ Å. This structure was originally solved using multiple isomorphous replacement and molecular replacement procedures (Zanotti *et al.*, 1992). The model of adipocyte lipid binding protein (A-LBP), obtained at 2.5 Å resolution, was used as search model for molecular replacement; it shares 64% amino acid identities with M-FABP. The rotation function in *MERLOT* (Fitzgerald, 1988) was used to orient the molecule using data between 15.0 and 4.0 Å resolution and a translation search was made by *X-PLOR* (Brünger, 1990) using 1351 reflections between 15 and 4.0 Å resolution.

LPH (*Lucina pectinata* hemoglobin I): $P2_1$, $a = 37.97$, $b = 38.39$, $c = 42.65$ Å, $\beta = 97.40°$. This structure has been solved by molecular replacement (Rizzi *et al.*, 1994) using as search molecule the molecular model of *A. limacina* myoglobin, whose amino acid sequence shares 25% identities with LPH. The program *AMoRe* (Navaza, 1994) was used throughout for both rotation- and translation-function determination using data between 10.0 and 3.0 Å resolution.

STM (sea turtle myoglobin): $P2_12_12_1$, $a = 37.5$, $b = 61.1$, $c = 75.2$ Å. The structure determination was promptly obtained by molecular replacement using the program *AMoRe* (Navaza, 1994). Sperm whale myoglobin was employed as search model, using data from 15.0 to 4.0 Å resolution range for both rotational and translational searches (Rizzi *et al.*, 1993; Nardini *et al.*, 1995). Sperm whale myoglobin shares 63% amino acid identities with STM.

XSD (*Xenophous leavis* superoxide dismutase): $P2_12_12_1$, $a = 73.33$, $b = 68.86$, $c = 59.73$ Å. Cu, Co bovine SOD has been used as a search model for the structure determination by means of molecular replacement. The program *AMoRe* was employed, using data between 15.0 and 4.0 Å resolution (Djinovic Carugo *et al.*, 1993). The amino acid sequence homology between the search model and XSD is ~50%.

6. The first applications of the Main formula

Before applying the Main formula to experimental data, we will briefly study its efficiency in the ideal situation characterized by the following protocol (protocol 1): (*a*) the structure factors $F_{\mathbf{h}}$ are calculated from the published crystal structure up to experimental resolution. In this case, the $|E|_M^2$'s used in (1) are devoid of experimental errors; (*b*) the model molecule coincides with the entire asymmetric unit of the protein test structure (then $N_f = 1$). The translation problem was simulated by rigidly translating all the symmetry-independent atoms by $\tau = [0.3\mathbf{a}, 0.3\mathbf{b}, 0.3\mathbf{c}]$ from their true positions. Such modified sites constitute the set of

Table 1. *Code name, space group and crystallochemical data for test structures*

Code name	Space group	Nref	RES (Å)
M-FABP[(a)]	$P2_12_12_1$	7595	2.14
LPH[(b)]	$P2_1$	17352	1.50
STM[(c)]	$P2_12_12_1$	9758	1.97
XSD[(d)]	$P2_12_12_1$	19056	2.01

References: (*a*) Zanotti *et al.* (1992); (*b*)Rizzi *et al.* (1994); (*c*) Nardini *et al.* (1995); (*d*) Djinovic Carugo *et al.* (1993).

Table 2. *Protocol 1: statistical calculations for triplet invariants estimated via equation (1) for M-FABP and LPH*

NR is the number of triplets having $Q >$ ARG, $\langle|\Delta\Phi|\rangle$ is the average error, % is the percentage of triplets with $|\Delta\Phi| < \pi/2$.

	M-FABP			LPH		
ARG	NR	%	$\langle\|\Delta\Phi\|\rangle$	NR	%	$\langle\|\Delta\Phi\|\rangle$
0.0	19843	83.4	48	8970	71.2	65
0.4	19588	83.7	48	8952	71.2	64
2.0	12537	89.1	40	8554	71.8	64
3.2	5509	91.3	37	7855	72.6	63
4.4	1743	92.1	35	6853	73.3	62
6.5	242	91.3	33	4623	72.5	63
15.0	9	100.0	9	672	67.3	68

positional vectors $\mathbf{u}_j$ by which the $g_{is}(\mathbf{h})$ functions are calculated. Protocol 1 (*i.e.* calculated structure-factor moduli and model molecule coincident with the protein molecule) generates a nonrealistic and overoptimistic situation but will enable the reader to evaluate the potentiality and the limits of the Main formula. For the sake of brevity, such a formula will be preliminarily applied to M-FABP and LPH only, which are the most difficult cases to solve. Our final tests will concern all four test structures.

The reflections of M-FABP and LPH were ranked in decreasing order of $|E_M|$. NLAR = 800 and NLAR = 1000 are the respective number of reflections among which the triplets invariants are found. In Table 2, a statistical check on the reliability of the Main formula is shown. N_r is the number of triplets having $Q >$ ARG and

$$\langle|\Delta\Phi|\rangle = \langle|\Phi_{\text{true}} - \Phi_{\text{est}}|\rangle$$

is the corresponding average of the absolute difference between the 'true' (corresponding to the published test structure) triplet phase and the triplet phase estimated *via* (1). % is the percentage of triplets for which $|\Delta\Phi|$ is smaller than $\pi/2$. Table 2 clearly shows that:

(*a*) the Main formula may overestimate the triplet reliability;

Table 3. *Protocol 2: statistical calculations for triplet invariants estimated via equation (1) for M-FABP and LPH*

NR is the number of triplets having $Q >$ ARG, $\langle|\Delta\Phi|\rangle$ is the average error, % is the percentage of triplets with $|\Delta\Phi| < \pi/2$.

	M-FABP			LPH		
ARG	NR	%	$\langle\|\Delta\Phi\|\rangle$	NR	%	$\langle\|\Delta\Phi\|\rangle$
0.0	24787	50.8	89	13409	50.4	89
0.4	24206	50.7	89	13354	50.4	89
2.0	12362	50.8	88	11949	50.4	89
3.2	4878	52.4	87	10067	50.4	89
4.4	1835	52.2	87	7835	50.5	89
6.5	346	53.5	87	4155	50.7	89

(*b*) the number of wrongly estimated triplets is remarkably large, in spite of the fact that protocol 1 creates the most favourable situation. The above outcome also suggests that, if a phasing procedure is applied *via* the above triplets to M-FABP and LPH data, the assigned phases should be characterized by large values of α but the solution of the translational problem would not be straightforward. That is exactly what we obtain when we apply the procedure described in §4. In particular: (i) for M-FABP the true solution is found with mean phase error equal to 40° (over NLAR = 800 reflections) but it is ranked in the fourth position by the figures of merit; (ii) for LPH no solution is found. In both cases, 200 trial solutions were explored.

Let us come to the case in which real experimental data (up to experimental protein resolution) and real model molecules in §5 are used (protocol 2). Table 3 is obtained. Comparison with Table 2 shows that: (*a*) the efficiency of (1) collapses, which is, according to the results obtained *via* protocol 1, mainly due to the lack of correlation between the model molecule and the test structure; (*b*) the overestimation of the triplet reliability increases and quite large Q values are often associated with wrong estimates. As a consequence, the correlation between the accuracy parameter Q and the triplet reliability tends to vanish. Accordingly, the straightforward application of the Main formula does not succeed: no solution is found among the 200 trial solutions.

An additional test was made: in order to increase the correlation between the model molecule and the structure under study, only reflections up to 4 Å resolution were involved in the calculations. The results can be summarized as follows: a much larger number of triplets was found among the 800 largest $|E_M|$ reflections, the overestimation of the triplets decreases but again no solution is found among 200 trial solutions. The above tests suggest that the application of direct methods for solving the translation problem at 4 Å resolution is practicable provided some modification to the Main formula is afforded.

Table 4. *Protocol 2: statistical behaviour of the M-FABP phases against ΔE after submission of the true phases to tangent refinement*

Protocol 3 is used. NR is the number of reflections having $\Delta E >$ ARG, $\langle|\Delta\Phi|\rangle$ is the average error.

$\Delta E > 0$			$\Delta E < 0$		
ARG	NR	$\langle\|\Delta\Phi\|\rangle$	ARG	NR	$\langle\|\Delta\Phi\|\rangle$
0.00	178	57	−2.20	620	72
0.10	139	55	−1.32	603	73
0.15	113	54	−0.88	540	73
0.25	81	53	−0.66	456	71
0.30	68	50	−0.44	324	69
0.55	30	66	−0.20	210	70

7. The new procedure

A special test may be applied for judging the efficiency of the Main formula: the correct phase values of the NLAR reflections are submitted to tangent refinement. Once convergence has been attained, the final phase values are expected to present large or small deviations from the starting values according to whether the average triplet reliability is high or low. Then the average phase error $\langle|\Delta\Phi|\rangle$ can be considered as the best result achievable by application of the tangent formula *via* a multisolution procedure. When such a test has been applied to M-FABP experimental data at experimental resolution (protocol 2), we obtained $\langle|\Delta\Phi|\rangle = 69^\circ$, which confirms the inefficiency of our procedure and of the Main estimates. The new phases were analysed against α, $|E_W|$ and $\Delta E = |E_W| - |E_M|$. We found that:

(*a*) α and $|\Delta\Phi|$ are inversely correlated: *e.g.* $|\Delta\Phi| = 56^\circ$ for the 159 phases with $\alpha > 83$, $|\Delta\Phi| = 59^\circ$ for the 374 phases with $\alpha > 55$. This correlation (encouraging but not sufficiently high) is due to the fact that % is larger than 0.5 for most of the Q values.

(*b*) $|E_W|$ and $|\Delta\Phi|$ are inversely correlated: *e.g.* $|\Delta\Phi| = 54^\circ$ for the 140 reflections with $|E_W| > 2.0$, $|\Delta\Phi| = 62^\circ$ for the 412 reflections with $|E_W| > 1.55$.

(*c*) ΔE and $|\Delta\Phi|$ are directly and strongly correlated (at least for reflections with $\Delta E > 0$). This was rather unexpected and will have important consequences on our strategy.

In Table 4, $\langle|\Delta\Phi|\rangle$ is given against ΔE, for reflections with $\Delta E > 0$ and $\Delta E < 0$, respectively. Only 178 reflections with $\Delta E > 0$ are among the NLAR reflections (the NREF reflections were ordered according to decreasing $|E_M|$ values). We note that reflections with $\Delta E > 0$ show deviations from the true values ($\langle|\Delta\Phi|\rangle = 57^\circ$) remarkably smaller than reflections for which $\Delta E < 0$ ($\langle|\Delta\Phi|\rangle = 72^\circ$). The above results suggest that it may be better to order the NREF reflections in decreasing order of $|E_W|$ and choose the NLAR reflections as those with the largest $|E_W|$ values.

The experimental results above described suggest the following procedure:

(i) A threshold is fixed for the resolution. It mostly depends on the similarity between model molecule and protein. The use of high-resolution data is advisable only in the case of high similarity.

(ii) Reflections are ordered according to ΔE: *e.g.* the first in the ordered set is the reflection with the largest positive value of ΔE, the last in the set is that with the largest negative value of ΔE.

(iii) A threshold TR ΔE is fixed: reflections with $\Delta E >$ TR ΔE are selected and, among them, NLAR reflections with $|E_W| >$ TR E_W are used for the triplet invariant search. TR ΔE and TR E_W are not critical values: in our tests, TR ΔE is usually between 0.0 and 0.4 and TR $E_W \geq 0.4$. To be more explicit, if TR $\Delta E = 0$ and TR $E_W = 0.4$, the NLAR reflections satisfy the conditions $\Delta E > 0$ and $E_W > 0.4$. The threshold values are chosen so as to select a sufficiently large number of reflections among which reliable triplets could be found. The most striking feature of the process is that even reflections with small values of E_W are expected to give rise to reliable triplets provided $\Delta E > 0$.

(iv) Reflections for which $|E_W|$ and $|E_M|$ are simultaneously very weak are selected for the psizero figure of merit (see §8).

(v) The triplets are estimated *via* the Von Mises formula

$$P(\Phi) \approx [2\pi I_0(G)]^{-1} \exp[G\cos(\Phi - \Theta)], \qquad (8)$$

where

$$G = p_1 G',$$

$$G' = 2p_2 |E_{W\mathbf{h}_1} E_{W\mathbf{h}_2} E_{W\mathbf{h}_3}| \times \left[\frac{Q'^2 + Q''^2}{\langle|F_{\mathbf{h}_1}|^2\rangle_M \langle|F_{\mathbf{h}_2}|^2\rangle_M \langle|F_{\mathbf{h}_3}|^2\rangle_M}\right]^{1/2}. \qquad (9)$$

p_1 is a weighting factor which limits the range of G to the interval (0, 6), to avoid triplets with too high values of the reliability factor G' dominating the phasing process. Q', Q'' and Θ are defined in (1). The reliability coefficient G' may be obtained from Q by replacing

$$\left\{\sum_N(\mathbf{h}_1)\sum_N(\mathbf{h}_2)\sum_N(\mathbf{h}_3)\Big/\langle|F_{\mathbf{h}_1}|^2\rangle_M\langle|F_{\mathbf{h}_2}|^2\rangle_M\langle|F_{\mathbf{h}_3}|^2\rangle_M\right\}^{1/2} \qquad (10)$$

by a unitary factor. Actually, (10) is often far from unity, so that (1) and (8) work quite differently. To give a practical example, let us consider a triplet for which

$$\varepsilon\sum_N(\mathbf{h}_i) < \langle|F_{\mathbf{h}_i}|^2\rangle_M, \quad \text{for } i = 1, 2, 3.$$

Then (10) is smaller than unity: assuming it equal to one increases the reliability of the triplets for which $|E_{Wi}| > |E_{Mi}|$. The smaller (10) is with respect to unity,

Table 5. *Protocol 1: statistical calculations for triplet invariants estimated via equation (9) for M-FABP and LPH (data up to experimental resolution)*

	M-FABP			LPH		
ARG	NR	%	$\langle\lvert\Delta\Phi\rvert\rangle$	NR	%	$\langle\lvert\Delta\Phi\rvert\rangle$
0.0	43480	77.6	57	19288	96.1	27
0.4	42683	78.0	56	19273	96.2	27
2.0	24281	86.6	45	18745	96.9	26
3.2	9936	92.6	36	17451	97.9	24
4.4	2968	94.1	32	13909	99.1	21
5.5	645	93.2	32	5116	99.7	18

Table 6. *Protocol 2: statistical calculations for triplet invariants estimated via equation (9) for M-FABP and LPH (data up to experimental resolution)*

	M-FABP			LPH		
ARG	NR	%	$\langle\lvert\Delta\Phi\rvert\rangle$	NR	%	$\langle\lvert\Delta\Phi\rvert\rangle$
0.0	47078	55.4	84	20125	51.5	88
0.4	44704	55.6	84	20071	51.5	88
2.0	11431	58.8	80	17851	51.9	88
3.2	2321	62.9	75	13243	52.2	88
4.4	343	72.0	67	57321	52.7	87
5.5	19	84.2	64	927	54.3	85

the larger the underestimate of the triplet provided by (1). On the contrary, if

$$\varepsilon \sum_N(\mathbf{h}_i) > \langle|F_{\mathbf{h}_i}|^2\rangle_M, \quad \text{for } i = 1, 2, 3,$$

assuming (10) equal to unity depresses the reliability of the triplets for which $|E_{Wi}| < |E_{Mi}|$.

p_2 is a weight factor that assumes different values according to the number of negative ΔE values. *I.e.* p_2 is 1 if all the three reflections contributing to the triplets have $\Delta E > 0$ [this is the case of equation (9)]; $p_2 = 0.8$, 0.6 or 0.4 when only one ΔE, two ΔE or three ΔE are negative.

Equation (8) has no theoretical basis: it has been suggested by our numerous applications. We will see in §9 that (8) not only provides better phase estimates but it also allows figures of merit (FOM's) to work efficiently.

8. The figures of merit

Figures of merit are essentially those proposed by Cascarano *et al.* (1992) but modifications were necessary to secure their successful use. The MABS figure of merit (Declercq *et al.*, 1979) is not actively used. The figure ALFCOMB, besides the differences $(\alpha - \langle\alpha\rangle)$, involves the standard deviation of α, σ_α. Since the model structure may be weakly correlated with the protein under study, the estimated variance may strongly underestimate the true variance. In order to take this effect into account, in our calculations we multiply σ_α^2 by 30. The traditional psizero figure of merit (Cochran & Douglas, 1957) has been modified by introducing suitable weights taking into account the prior information:

$$\mathrm{PSI}(0) = \sum_{\mathbf{h}} w_{\mathbf{h}}\alpha'_{\mathbf{h}} \Big/ \sum_{\mathbf{h}} v_{\mathbf{h}}^{1/2},$$

where

$$\alpha'_{\mathbf{h}} = \left| \sum_j w_j E_{W\mathbf{k}_j} E_{W\mathbf{h}-\mathbf{k}_j} \right|,$$

$$v_{\mathbf{h}} = \sum_j w_j^2 |E_{W\mathbf{k}_j} E_{w\mathbf{h}-\mathbf{k}_j}|^2,$$

$$w_j = \left\{ (Q_j'^2 + Q_j''^2) / \langle|F_{\mathbf{k}_j}|^2\rangle_M \langle|F_{\mathbf{h}-\mathbf{k}_j}|^2\rangle_M \right\}^{1/2},$$

$$Q_j' = \Re\left\{ \sum_{i=1}^{N_f} \sum_{s=1}^{m} g_{is}(\mathbf{k}_j) g_{is}(\mathbf{h}-\mathbf{k}_j) + \sum_{2q}(\mathbf{k}_j, \mathbf{h}-\mathbf{k}_j) \right\},$$

$$\sum_{2q}(\mathbf{k}_j, \mathbf{h}-\mathbf{k}_j) = \sum_{j=1}^{q} f_j(\mathbf{k}_j) f_j(\mathbf{h}-\mathbf{k}_j),$$

$$Q_j'' = \Im\left\{ \sum_{i=1}^{N_f} \sum_{s=1}^{m} g_{is}(\mathbf{k}_j) g_{is}(\mathbf{h}-\mathbf{k}_j) \right\}.$$

ALFCOMB and PSI(0) are combined in a suitable CFOM. We will see in the next section that CFOM is able to discriminate the correct solution in all the test cases.

9. Experimental applications

We first apply (8) to control its efficiency in ideal conditions (protocol 1). 800 reflections for M-FABP and 1000 reflections for LPH are selected according to the procedure described in §7. The corresponding triplet statistics are shown in Table 5. We observe: (*a*) the NLAR reflections selected by the new procedure are highly connected by the triplet network – NR values in Table 5 are double those in Table 2; (*b*) triplets are ranked by (8) better than by (1). The improvement is dramatic for LPH, less evident but non-negligible for M-FABP. For example, for this last structure, for the 1743 triplets with $Q > 4.4$ the percentage of the correctly estimated cosines is 92.1, while for the 2968 triplets with $G > 4.4$ the percentage is 94.1.

Let us now apply (8) to real data up to experimental resolution (protocol 2): again NLAR = 800 and 1000 for M-FABP and LPH, respectively. The triplet statistics are summarized in Table 6 and may be usefully compared with results in Table 3. The improvement in terms of the number of triplets and the mean phase error for M-FABP is evident and does not deserve further discussion. Less clear is the improvement for LPH in terms of mean phase error. We will see below that no solution is obtained at experimental resolution for such

a structure, while a solution is found if data up to 4 Å resolution are used. For reader usefulness in Table 7, we show the corresponding triplet statistics.

The procedure described in §8 is applied to the four test structures. In order to check if the translation problem was correctly solved, we calculated the correlation factor CORR between the electron density ρ calculated from the phases assigned by our direct-methods procedure and the 'true' map ρ_{true} (corresponding to the published phases):

$$\text{CORR} = \frac{\langle \rho \rho_{\text{true}} \rangle - \langle \rho \rangle \langle \rho_{\text{true}} \rangle}{(\langle \rho^2 \rangle - \langle \rho \rangle^2)^{1/2} (\langle \rho_{\text{true}}^2 \rangle - \langle \rho_{\text{true}} \rangle^2)^{1/2}}.$$

200 trials per structure are always performed, starting from random sets of phases. The results may be summarized as follows:

M-FABP: 3 Å resolution data are used: the highest value of CFOM (= 0.393) recognizes the correct solution. NLAR = 641 reflections are first phased with a mean phase error of 38°. The phase-expansion process leads to 7560 phased reflections with mean phase error of 72°. The value of CORR is 0.45. The electron-density map calculated with all the phased reflections is of good overall quality, allowing straightforward model-building interpretation throughout the polypeptide chain (see Fig. 1).

LPH: No solution is found with 3 Å resolution data. A 4 Å cutoff is then used: the highest value of CFOM (= 0.558) singles out the correct solution. NLAR = 357 reflections are first phased, with a mean error of 49°. The phase-expansion process leads to a total of 7085 phased reflections with mean phase error of 80°. CORR is equal to 0.29. Inspection of the electron-density map, calculated with the 7085 extended phases and observed structure factors, shows regions that can be easily interpreted in terms of an atomic model compatible with the previously determined structure of the protein. In particular, regions of immediate interpretability are the α-helical segments surrounding the heme group, which is well defined (see Fig. 2). On the other hand, less clear electron density is obtained for regions of the protein structure further away from the heme, for which in the absence of additional information an atomic model cannot be fitted.

Table 7. *Protocol 2: statistical calculations for triplet invariants estimated via equation (9) for LPH (data up to 4 Å resolution, NLAR = 396)*

LPH			
ARG	NR	%	$\langle \lvert \Delta\Phi \rvert \rangle$
0.0	17747	55.6	83
0.4	13651	56.0	83
2.0	4099	60.2	78
3.2	1319	60.9	77
4.4	346	61.8	77
5.5	53	64.2	73

STM: 3 Å resolution data are used: the highest value of CFOM (= 0.437) singles out the correct solution. NLAR = 621 are first phased with a mean phase error of 20°. The phase-expansion process leads to a total of 9753 reflections with a mean phase error of 53°. CORR = 0.73. The corresponding phases allowed the calculation of an electron-density map which could be easily interpreted in terms of the final molecular model of the protein, the quality of the electron density being constant throughout the asymmetric unit (see Fig. 3).

XSD: 3 Å resolution data are used: the highest value of CFOM (= 0.290) singles out the correct solution. NLAR = 680 reflections are first phased with mean

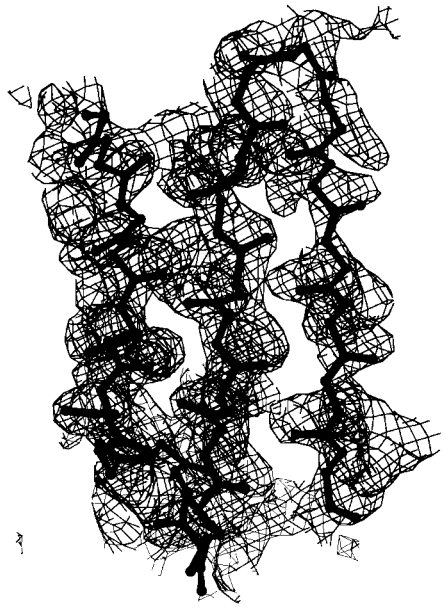

Fig. 1. M-FABP: a region showing the calculated electron-density map for three β-strands contoured at the 1.0σ level.

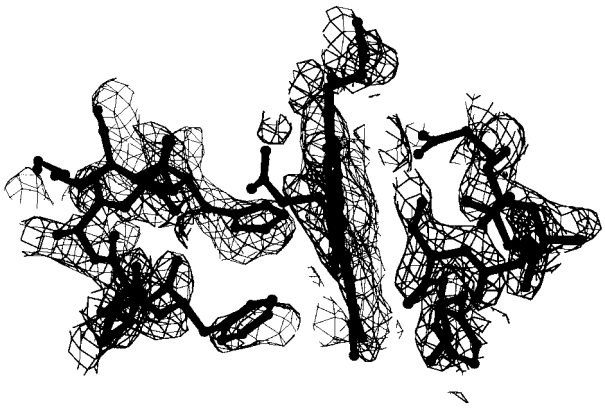

Fig. 2. LPH: electron density calculated for the translated molecule showing the heme group (edge on in the picture) and portions of two surrounding helices.

phase error of 27°. The phase-expansion process leads to a total of 19016 reflections with mean phase error of 66°: CORR = 0.57. Fig. 4 shows a region comprising the Cu^{2+} and Zn^{2+} ions in the enzyme active site, calculated with the 19016 phased reflections. The dimeric enzyme model fits very nicely the electron density throughout the asymmetric unit and model building, in the absence of a molecular model, could have been easily achieved.

10. Final remarks

This paper describes the first successful application of direct methods to the translation problem in macromolecular crystallography. The pioneering contribution by Main (1976) plays a central role: his formula (1) was designed to take into account the prior information available when a model fragment is correctly oriented. While (1) proved to be a useful tool for the small-molecule field, some modifications were necessary to improve its efficiency for macromolecules. Our experimental tests led us to suggest the new formula (8), which is more able to overcome the problems usually met in macromolecular crystallography. It may be worthwhile observing that (8) [as well as (1)] exploits a type of prior information that is complementary to that used by most of the techniques currently used in molecular replacement techniques. For example, several translation functions are based on Patterson convolutions (Buerger, 1959; Hoppe, 1957; Huber, 1965):

$$T(\mathbf{t}) = \int P_0(\mathbf{u}) P_p(\mathbf{u}, \mathbf{t})\, \mathrm{d}\mathbf{u},$$

where $P_0(\mathbf{u})$ is the observed Patterson at point $\mathbf{u}$, $P_p(\mathbf{u}, \mathbf{t})$ is the Patterson calculated from all the molecular fragments symmetry related to the model molecule (inter- and intramolecular vectors included). Better results are usually obtained when self-Patterson peaks are subtracted from $P_0(\mathbf{u})$ and $P_p(\mathbf{u}, \mathbf{t})$. Then,

$$T(\mathbf{t}) = \int \left[P_0(\mathbf{u}) - \sum_{s=1}^{m} P_s(\mathbf{u}) \right] \left[P_p(\mathbf{u}, \mathbf{t}) - \sum_{s=1}^{m} P_s(\mathbf{u}) \right] \mathrm{d}\mathbf{u},$$

where $P_s(\mathbf{u})$ is the calculated Patterson function for the sth fragment symmetry equivalent to the input fragment. The Fourier transform of $T(\mathbf{t})$ gives the so-called correlation functions (Rossmann & Blow, 1962; Tollin, 1966; Crowther & Blow, 1967; Beurskens, 1981). If we compare (1) with $T(\mathbf{t})$, it becomes clear that: (*a*) equation (1) exploits, as prior information, the intramolecular vectors of the model structure only; (*b*) Patterson convolution methods are mainly based on intermolecular vectors.

The procedure we proposed proved highly efficient in the case of very high sequence homology between the search model and the unknown protein, providing electron-density maps which, in the cases of STM (an α-helical protein) and XSD (an antiparallel β-barrel structure), can unambiguously be interpreted. On the other hand, despite the translational search providing the correct solution, in the case of LPH the electron density calculated did not allow the complete modelling of the protein molecule. In all cases, we noticed that for several residues spurious density, clearly reflecting the model information present in the search molecule adopted, was visible at the 1.0σ contour level. On the other hand, the calculated electron densities did not allow the fitting of molecular models beyond the limits of the search molecules adopted (*i.e.* no side-chain density for trimmed or omitted residues was present or interpretable). Nevertheless, in all the four cases here presented, the calculated electron density could locate properly the unknown molecules, using an approach that

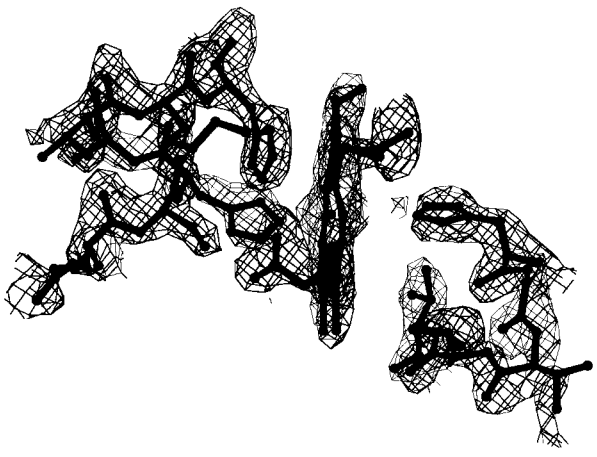

Fig. 3. STM: electron density displaying approximately the same structural region selected for the homologous LPH protein in Fig. 2.

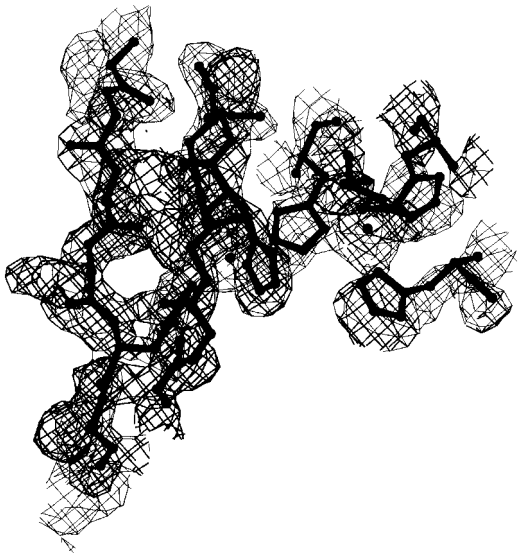

Fig. 4. XSD: a region of the XSD electron-density map, after proper translation, comprising the Cu^{2+} and Zn^{2+} ions (shown as isolated atoms) in the enzyme active site.

is new in macromolecular crystallography. We are aware that, with respect to conventional molecular replacement techniques, the new procedure here applied may be slower, since it does not provide atomic coordinates for the translated model but only phases for the calculation of the respective electron-density maps. However, the method here presented maybe improved in several aspects:

(*a*) *Via* a better use of the vibrational atomic parameters of the search model. In our procedure, an overall isotropic thermal factor, equal to that calculated for the protein by a Wilson plot, is associated with all the atoms of the model.

(*b*) Once the protein phases are available, the translation vector may be straightforwardly found *a posteriori* by the fast-Fourier-transform-based algorithm recently proposed by Lunin & Lunina (1996). Atomic coordinates for the translated model will then be available.

(*c*) Quartet invariant estimates exploiting intramolecular vectors of the model structure might be associated with triplet invariants in the phasing procedure. A probabilistic formula so designed is described in the second paper of this series (Giacovazzo *et al.*, 1997).

(*d*) A solvent-flattening procedure may be used to improve the quality of the phases provided by our translation procedure [see Giacovazzo & Siliqi (1997) for an effective solvent-flattening procedure applied to direct-methods phases].

All these aspects will be faced in the next paper of this series.

Part of this work has been supported by the European Union TMR grant No. CT94-0690 'Advanced Methods for the Crystallography of Biological Macromolecules'.

References

Altomare, A., Cascarano, G., Giacovazzo, C., Guagliardi, A., Burla, M. C., Polidori, G. & Camalli, M. (1994). *J. Appl. Cryst.* **27**, 435.

Baggio, R., Woolfson, M. M., Declercq, J. P. & Germain, G. (1978). *Acta Cryst.* A**34**, 883–892.

Beurskens, P. T. (1981). *Acta Cryst.* A**37**, 426–430.

Beurskens, P. T., Gould, R. O., Bruins Slot, H. J. & Bossman, W. P. (1987). *Z. Kristallogr.* **179**, 127–159.

Brünger, A. T. (1990). *XPLOR* version 2.1 manual. *A System for Crystallography and NMR.* Yale University Press, New Haven, CT, USA.

Buerger, M. J. (1959). *Vector Space.* New York: Wiley.

Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1992). *Acta Cryst.* A**48**, 859–865.

Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.

Cochran, W. & Douglas, A. S. (1957). *Proc. R. Soc. London Ser. A*, **243**, 281.

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

Declercq, J. P., Germain, G. & Woolfson, M. M. (1979). *Acta Cryst.* A**35**, 622–626.

Djinovic Carugo, K., Collyer, C., Coda, A., Carrí, M. T., Battistoni, A., Bottaro, G., Polticelli, F., Desideri, A. & Bolognesi, M. (1993). *Biochem. Biophys. Res. Commun.* **194**, 1008–1011.

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

Giacovazzo, C., Cascarano, G. & Zheng, C. (1988). *Acta Cryst.* A**44**, 45–51.

Giacovazzo, C., Manna, L. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 799–806.

Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.

Giacovazzo, C., Siliqi, D., Gonzalez Platas, J., Hecht, H. J., Zanotti, G. & York, B. (1996). *Acta Cryst.* D**52**, 813–825.

Hoppe, W. (1957). *Acta Cryst.* **10**, 750–751.

Huber, R. (1965). *Acta Cryst.* **19**, 353–356.

Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* A**34**, 863–870.

Langs, D. A., Guo, D. & Hauptman, H. A. (1995). *Acta Cryst.* D**51**, 1020–1024.

Lunin, V. Yu. & Lunina, N. L. (1996). *Acta Cryst.* A**52**, 365–368.

Main, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, pp. 97–105. Copenhagen: Munksgaard.

Nardini, M., Tarricone, C., Rizzi, M., Lania, A., Desideri, G., De Sanctis, G., Coletta, M., Petruzzelli, R., Ascenzi, P., Coda, A. & Bolognesi, M. (1995). *J. Mol. Biol.* **247**, 459–465.

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

Rizzi, M., Ascenzi, P., Coda, A., Brunori, M., Bolognesi, M. (1993). *Rend. Fis. Acc. Lincei*, **4**, 65–73.

Rizzi, M., Wittemberg, J. B., Coda, A., Fasano, M., Ascenzi, P. & Bolognesi, M. (1994). *J. Mol. Biol.* **244**, 86–99.

Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.

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

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

Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.