Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 21 August 1999
Accepted 22 October 1999

Upgrading the twin variables algorithm for large structures

K. Bethanis,[a] P. Tzamalis,[a] A. Hountas,[a]* A. F Mishnev[b] and G. Tsoucaris[c]

[a]Physics and Meteorology Laboratory, Agricultural University of Athens, 75 Iera Odos, Votanikos, Athens 118-55, Greece, [b]Institute of Organic Synthesis, Latvian Academy of Sciences, 1006 Riga, Latvia, and [c]Laboratoire de Physique, Tour B, Centre Pharmaceutique, Université Paris Sud, 92290 Châtenay-Malabry, France. Correspondence e-mail: gphy2xoa@auadec.aua.gr

Phase extension from lower to higher resolution by using an upgraded *TWIN* variables algorithm [Hountas & Tsoucaris (1995). *Acta Cryst.* A**51**, 754–763] in protein molecules with close to 1000 non-H atoms is presented. Three points of this procedure are of particular interest. (i) The use of a set of auxiliary variables providing a satisfactory fit for many kinds of constraints: the new algorithm works efficiently despite the extreme 'dilution' of very limited initial phase information into a much larger set of auxiliary variables. (ii) The extension of this auxiliary variables set beyond the resolution of the observed data, which enhances the phase extension in a so-called 'super-resolution' sphere. (iii) The use of the crystallographic symmetry as a new figure of merit and as a reliable test for the correctness of the phase-extension process allows an efficient screening.

1. Notation and definitions

AMIN: cut-off value for acceptance for each reflection based on the modulus of the calculated E
MPE: mean phase error
M_{mod}: moduli minimization function
M_{sf}: structure-factor minimization function
S-FOM: crystallographic 'symmetry figure of merit' – a measure of the inconsistency among the calculated phases and magnitudes for symmetry-related reflections
S_MPE: symmetry mean phase error – mean phase error among symmetry-related reflections
S_R_{mod}: symmetry mean modulus error
Ψ_S_MPE: Ψ symmetry mean phase error.

2. Introduction

An important stage in macromolecular crystallography is that of phase extension and refinement when initial phase estimates are available from isomorphous replacement or anomalous scattering. On the other hand, it has been shown recently that direct methods are able to provide *ab initio* an approximate or partial solution for small protein structures, with up to 1000 atoms in the asymmetric unit (Smith *et al.*, 1996; Shaefer *et al.*, 1998). In most cases, it is necessary to extend the phases either from lower to higher resolution or within the same resolution range. For phase improvement and extension in small proteins, several methods have been used, such as direct-space FFT convolution (Barret & Zwick, 1971), Sayre's equation (Sayre, 1974), the tangent formula (Blundell *et al.*, 1981), the maximum determinant rule (de Rango *et al.*, 1985), the Sayre-equation tangent formula (Woolfson & Yao, 1988) and others. One of the most successful techniques of phase extension and refinement uses density modification (Podjarny *et al.*, 1987). In its various forms, it applies to the density constraints such as positivity, atomicity, boundedness, solvent flatness, connectivity and non-crystallographic symmetry. A recent addition to density modification is histogram matching which imposes the correct density histogram on the map (Zhang & Main, 1990). Another approach has been developed on the basis of maximum entropy and likelihood concepts (Roversi *et al.*, 1998).

The present paper describes further developments of the twin variable method (Hountas & Tsoucaris, 1995; hereafter called H–T) including upgrading of the algorithm for protein phase extension and refinement. The upgrading comprises (*a*) an increase in the size of the molecules from 200 to 1000 independent non-H atoms and (*b*) achievement of considerable extension in the resolution of the phases determined by the present algorithm, for instance from initial given phases at 3.5 Å to final phases at 1.2 Å.

Two novelties are particularly relevant in the context of direct methods, respectively, in steps (*a*) and (*b*) of the algorithm (Fig. 1), as follows.

(i) The use from the very beginning of a very large set of auxiliary variables defined on reciprocal-lattice vectors that can now be located beyond the limiting resolution of the observed structure factors. Despite the extreme dilution of very limited initial phase information into this large set of auxiliary variables, the algorithm allows phase refinement and phase extension both within the observed reciprocal sphere and the so-called super-resolution shell.

(ii) The use of the crystallographic symmetry as a new figure of merit (S-FOM) and as a reliable test for the correctness of the phase-extension process. One of the most serious problems in the application of direct methods for large structures is to find a reliable figure of merit. The introduction of S-FOM consists of testing the phase-extension and refinement algorithm by deliberately sacrificing the space-group symmetry in the starting set, then using its re-appearance as a criterion for correctness.

Preliminary results have been given in two communications (Tsoucaris *et al.*, 1996; Bethanis *et al.*, 1997).

3. Upgrading the *TWIN* algorithm

The twin variable concept consists of the use of a set of auxiliary complex variables $\Psi_{\mathbf{K}}$ associated with reciprocal-lattice vectors **K**. The Ψ set is related to the normalized structure factors (s.f.) by means of the following equations:

$$E_{\mathbf{H}} = \sum_{\mathbf{K}} \Psi_{\mathbf{K}}(\Psi_{\mathbf{K-H}})^* \overset{\text{FT}}{\Longleftrightarrow} \rho(r) = \left|\psi(r)\right|^2, \qquad (1)$$

$$\Psi_{\mathbf{K}} = \sum_{\mathbf{H}} E_{\mathbf{H}}\Psi_{\mathbf{K-H}}. \qquad (2)$$

Equation (2) is the so-called regression equation of standard probability theory as shown in H–T. The couple ($E_{\mathbf{H}}$, $\Psi_{\mathbf{H}}$) are called twin variables.

The Ψ variables alone control the whole procedure; they are allowed to change both in modulus and in phase (or real and imaginary parts) throughout the procedure. For instance, we write here explicitly one of the constraint functions considered in §4 [equation (5) in Fig. 1] to be minimized with respect to the real and the imaginary parts of all auxiliary variables Ψ,

$$\mathrm{M}_{\mathrm{mod}} = \sum_{\mathbf{H}}\left(\left|E_{\mathbf{H}}^{\mathrm{obs}}\right| - \left|\sum_{\mathbf{K}} \Psi_{\mathbf{K}}(\Psi_{\mathbf{K-H}})^*\right|\right)^2.$$

These optimal values of Ψ will then be introduced into (1) in order to obtain the phases of E.

An important feature of the *TWIN* algorithm is the essence of the fitting process exemplified in the above constraint. This fitting stems from the defining equation (1) and therefore it can be achieved with any desired accuracy. In practice, this accuracy should be somewhat better than the expected error on $|E_{\mathbf{H}}^{\mathrm{obs}}|$. Thus, the algorithm aims at determining the phases of E through a very large Ψ set, by satisfying a battery of constraints, though 'remaining' in the subspace of Ψ that fulfils the above constraint with the desired accuracy. Note that $\mathrm{M}_{\mathrm{mod}}$ can be readily reduced to practically zero but this clearly has no physical meaning and no practical use. This fact, however, greatly emphasizes how easy it is to construct (even positive) density functions whose Fourier coefficients have the same moduli as a given set of E^{obs}. This important remark calls for research into new criteria to help in discriminating the 'correct' solution; such a criterion will be presented in §6.

We summarize below the three steps of the *TWIN* algorithm (Fig. 1). We will illustrate the description using data from row 3 in Table 2. The initial input comprises the following.

(i) The whole set of 9772 unique observed moduli larger than a cut-off value $E_{\min} = 1.22$ at resolution 1.0 Å.

(ii) A small subset of 268 unique (symmetry-independent) initially phased $E_{\mathbf{H}}$ at resolution 3.6 Å.

An important feature of the algorithm is the use from the very beginning of a very large auxiliary Ψ set (20584 in the present calculations). In most of the present work, the Miller indices of the Ψ set are taken to be identical to those of the observed E. However, in a preliminary attempt to achieve a so-called 'super-resolution', the Ψ set is extended beyond the resolution of the observed E (§5).

It is important to emphasize that the complex Ψ variables are not restricted to satisfy either the Friedel law or the crystallographic space-group symmetry. However, in the present calculations, Ψ do satisfy the Friedel condition but not the symmetry conditions. Note that the latter enabled us to develop a new test based on symmetry (see §6). Thus, for a Ψ set with indices identical to these of the observed E ($P2_1$ structure with 10292 unique E), we have now 20584 Ψ to be

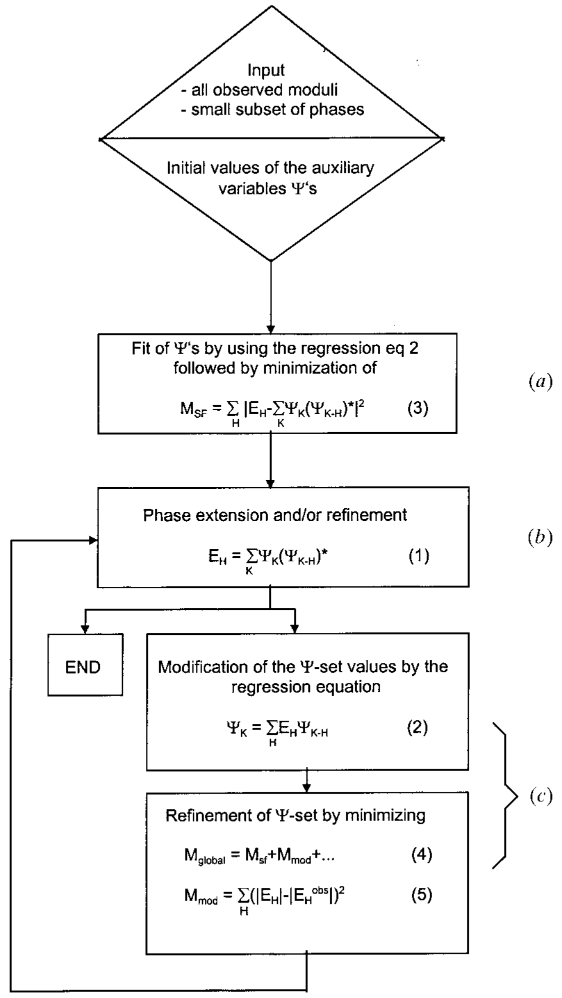
Input
- all observed moduli
- small subset of phases
Initial values of the auxiliary variables Ψ's
Fit of Ψ's by using the regression eq 2 followed by minimization of
$M_{SF} = \sum_H |E_H - \sum_K \Psi_K(\Psi_{K-H})^*|^2$ (3)
(a)
Phase extension and/or refinement
$E_H = \sum_K \Psi_K(\Psi_{K-H})^*$ (1)
(b)
END
Modification of the Ψ-set values by the regression equation
$\Psi_K = \sum_H E_H \Psi_{K-H}$ (2)
Refinement of Ψ-set by minimizing
$M_{global} = M_{sf} + M_{mod} + \ldots$ (4)
$M_{mod} = \sum_H (|E_H| - |E_H^{obs}|)^2$ (5)
(c)

Figure 1
Flow chart of the *TWIN* algorithm.

used as independent auxiliary variables. Their initial values have been chosen as follows.

(i) For a small subset of 268 Ψ, and the 232 Ψ symmetry-related by the 2_1 axis, we set their values equal to the values of the corresponding initially phased E (*i.e.* same indices, moduli and phases). However, this number of initially phased Ψ is not critical, unlike the number of initially phased E, and we have shown that the final results are only slightly impaired when this number is reduced to only 50.

(ii) For the remaining large subset of Ψ (*i.e.* 20084 in the typical example), we introduce arbitrary values in modulus and phase. For the initial phases, we use the *MS* random-number generation subroutine. For the initial moduli, the same subroutine could be used but in the present paper we simply set all moduli $|\Psi_{\mathbf{K}}|$ to an arbitrary value of 1.5.

The algorithm comprises three steps.

(*a*) *Preliminary transfer of phase information from the initially phased E set to the Ψ set.* The initially phased E set is to be introduced into (2) and in the following minimization function where it is kept constant throughout this step,

$$\mathrm{M_{sf}} = \sum_{\mathbf{H}} \left| E_{\mathbf{H}} - \sum_{\mathbf{K}} \Psi_{\mathbf{K}} (\Psi_{\mathbf{K-H}})^* \right|^2. \tag{3}$$

The first substep is the calculation of the whole set of 20584 Ψ by using the regression equation (2). Then the values of all Ψ are varied (in modulus and phase) so that their final values minimize the function $\mathrm{M_{sf}}$ by a steepest descent algorithm. Thus, the initial phase information contained in the unique 268 E is now capitalized into the much larger Ψ set (20584 Ψ). It is important to emphasize the 'extreme dilution' of the initial limited phase information into the large set of the auxiliary variables Ψ. This dilution is particularly striking in its lack of fulfillment of the space-group symmetry constraints (§6).

(*b*) *Stepwise phase extension by transfer of information from the Ψ set to the E set. New tests for accepting calculated phases of E.* The Ψ as determined in step (*a*) are introduced into (1) in order to determine the phases of new E (so-called extended phases). We achieve thus an inverse transfer of information, *i.e.* from the whole auxiliary Ψ set to unphased E. Thus, the initial phase information is now meaningfully transmitted through steps (*a*) and (*b*) into the phases of new E, *i.e.* beyond the initial known set. However, among the new E, very few are accepted in the subsequent iterative calculations, especially in the first iterations. Indeed, at this point it is very important to detect the most reliably phased E which will be in turn introduced, along with the 268 initially phased E, into step (*c*).

This is achieved through selection tests, such as the classical cut-off based on the modulus of the calculated E (AMIN test) and the new S-FOM related to symmetry (§6). We have thoroughly examined an optimal dependence of the AMIN cut-off value on the sequence number in the iteration process. This is a critical test especially in the first iterations. Thus, the AMIN value is varied gradually over 200 iterations from 1.8 to 0.8 (scale of normalized E) and a precise optimal AMIN/iteration table has been established. Fig. 2 shows the increase of the number of 'accepted' extended phases and the parallel decrease of MPE as a function of the AMIN value, itself determined by the iteration number. A similar technique could be used for the acceptance of contributors $\Psi_{\mathbf{K}}(\Psi_{\mathbf{K-H}})^*$ in the convolution equation (1), depending here on the cut-off value of TMIN = $|(E_{\mathbf{H}})^* \Psi_{\mathbf{K}} (\Psi_{\mathbf{K-H}})^*|$.

(*c*) *Phase refinement via a set of minimization functions of Ψ.* The values of the Ψ set are further varied so as to satisfy various constraints. This step comprises two parts.

(i) *The regression equation.* The phase extension and refinement procedure is greatly accelerated by using the regression equation (2) as a preliminary substep to the subsequent least-squares procedure. This equation directly generates modified values of $\Psi_{\mathbf{K}}$ as a function of the actual values of phased E and actual values of all other Ψ. It is to be noted that equation (2) is an approximation to the determinantal equation given by equation (1.5) in H–T. One can expect that the use of the complete determinantal equation will further improve the results.

(ii) *The global minimization function.* The Ψ set is finally varied so as to best satisfy the minimum condition of the global function,

$$\mathrm{M_{global}} = \mathrm{M_{sf}} + \mathrm{M_{mod}} + \ldots. \tag{4}$$

This function represents all constraints we wish to apply and, of course, includes $\mathrm{M_{sf}}$ and $\mathrm{M_{mod}}$,

$$\mathrm{M_{mod}} = \sum_{\mathbf{H}} \left(|E_{\mathbf{H}}| - |E_{\mathbf{H}}^{\mathrm{obs}}| \right)^2. \tag{5}$$

Moreover, we have provided optimization schemes for combining the two main mathematical tools: (i) the regression equation for storing the new information into the Ψ set and (ii) the least-squares process for a sum of explicit minimization functions. In the present calculations the programme has performed one regression cycle followed by nine least-squares cycles.

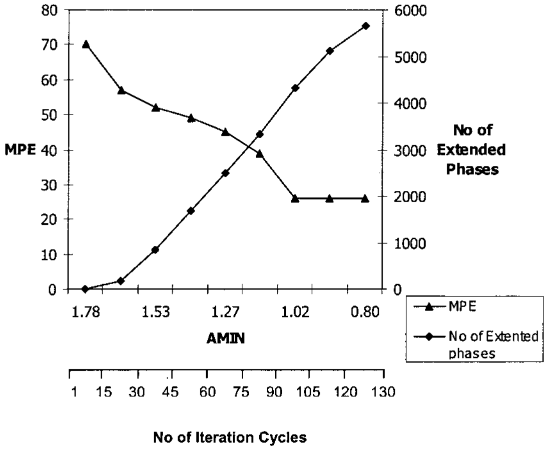

Figure 2
The value of AMIN (scale of normalized E) as a cut-off test for acceptance of new phases is determined by the iteration cycle. The large value of AMIN at the beginning is justified by the particular necessity for small phase errors at the early stages. The figure has been plotted with values that correspond to row 7 of Table 2.

Table 1
Names and chemical, unit-cell and symmetry data for protein structures.

Structure	Full name	Unit cell (Å, °)	Space group	No. of atoms in asymmetric unit cell	References
Rnase Ap1	Ribonuclease Ap1 of *Aspergillus pallidus*	a = 32.01, b = 49.76, c = 30.67, α = 90.0, β = 115.83, γ = 90.0	$P2_1$	890	Bezborodova *et al.* (1988)
1BKR	Calponin homology (ch) domain from human beta-spectrin	a = 31.65, b = 53.95, c = 32.35, α = 90.0, β = 105.48, γ = 90.0	$P2_1$	1095	Banuelos *et al.* (1998)†
1TMY	Chey from *Thermotoga maritima*	a = 32.04, b = 53.95, c = 34.16, α = 90.0, β = 95.56, γ = 90.0	$P2_1$	928	Usher *et al.* (1997)†

† Data were retrieved from the Protein Data Bank.

In addition, other *a priori* information can be included such as known solvent regions. A remarkable fact is that such a constraint in direct space can also be controlled by the Ψ set. This is achieved by minimizing the square of the electron density in the solvent region. It is to be noted that in H–T the minimization pertained to the ρ function, constrained to be non-negative by equation (4.14) of H–T

$$\mathrm{M}_{\rm solv} = \int\limits_{\rm solvent} \rho(r)\,\mathrm{d}v. \qquad (6)$$

In our present work, it has been shown that minimizing the following function is more efficient:

$$\mathrm{M}_{\rm solv} = \int\limits_{\rm solvent} \left[\rho(r) - \bar{\rho}_{\rm solvent}(r)\right]^2 \mathrm{d}v, \qquad (6a)$$

where $\bar{\rho}_{\rm solvent}(r)$ is the average density in the solvent region.

Other minimization functions have been the object of preliminary tests but they are not included in the present applications [non-observed reflections, negative quartets, the Cochran integral of ρ^3, mixed triplets from equation (4.5) of H–T].

The *TWIN* algorithm will proceed by iterating steps (b) and (c) until the complete set of reflections is phased.

4. Phase refinement and extension from low to high resolution in protein structures

The first step was to examine if the set of above equations (1)–(5) are valid for larger structures and for resolutions considerably lower than the atomic resolution. To achieve these goals, the *TWIN* algorithm has been upgraded and adapted to protein structures.

We have used both real and simulated data ('ideal', *i.e.* error free) for three small proteins (Table 1). For brevity, we use the following notations: ideal/ideal – moduli and phases calculated from refined atomic coordinates; real/ideal – observed E moduli obtained by subroutine *NORMAL* of program *MULTAN*88 (Debaerdemaeker *et al.*, 1988) and calculated phases; real/real – observed moduli and phases corrupted by imposition of random errors, generated by a wavy function, on the true values.

For the sake of detailed comparisons, we have taken practically the same number of ~260 initially phased E in all present calculations. This small number is very interesting in the frame of minimum information towards *ab initio* determination. The following points arise from the results shown in Table 2.

(i) In all ideal/ideal calculations, except that of no. 9, the final MPE (10–24°) is considered as satisfactory. However, by applying the super-resolution procedure for no. 9, the final MPE is again satisfactory. Outputs nos. 1, 2 and 3 show that the final MPE is roughly the same for a resolution range 2.5–3.6 Å.

(ii) For real/ideal calculations, we have two satisfactory outputs, nos. 4 and 6, whereas MPE is 78° for no. 11, at 1.9 Å resolution. Here again, the super-resolution procedure is likely to be efficient. The electron-density maps, generated using program O (Jones & Kjeldgaard, 1995), relative to the residues A33–A37 for no. 6 confirm that all atoms are correctly resolved after extension (Fig. 3).

(iii) For real/real calculations (nos. 5, 7 and 8), the MPE critically increases to 50–60° for an initial average error above ~20°. However, it is important to note that this critical value is related to the very small number of initially phased reflections (~250 reflections for ~1000 atoms) and that this work aimed at pushing the method to its extreme limit of initial phase information.

All results listed in Table 2 are considered as test cases for the present upgraded algorithm. However, current calculations, not reported here, have shown that the acceptable initial error considerably increases by increasing the number of phased reflections. This opens up the possibility of applying the algorithm in cases where a larger initial set of E is approximately phased with multiple isomorphous replacement (MIR) or anomalous dispersion.

5. Super-resolution calculations

'Super-resolution' is a term used to describe the fact that maximum-entropy methods can yield functions of resolution higher than that which corresponds to the band limits of the observed data (Collins, 1982). The fundamental equation (1) provides a very simple way to calculate phases beyond the observed data resolution.

The super-resolution effect can be better seen in comparison with the less good result of phase extension in calculation no. 9 of Table 2. It is to be noted that the resolution of the observed moduli is lower for nos. 9–11 than for all others.

Table 2
Summary of the results of phase extension from low to high resolution including the symmetry test.

In all calculations, the resolution of the extended set (columns noted Ext. set) is identical to that of the observed $|E|$, except for no. 10 pertaining to the super-resolution phase extension from 2 Å (observed moduli) to 1.5 Å. Symmetry MPE and R-factor calculations have not been performed for nos. 2 and 3.

			Resolution (Å)		No. of reflections			MPE (°)		R_{mod} factor	S_MPE	S_R_{mod} factor
Data set		Kind of data	Initial phased set	Ext. set	Initial phased set	Ext. set	No. of Ψ's	Initial phased set	Ext. set	Ext. set	Ext. set	Ext. set
1	Rnase Ap1	Ideal/ideal	2.5	1.0	268	4096	9257	0	15	7	1	1.1
2		Ideal/ideal	3.0	1.0	268	5833	11880	0	13	8	–	–
3		Ideal/ideal	3.6	1.0	268	9772	20584	0	10	10	–	–
4		Real/ideal	2.5	1.17	268	4455	12842	0	36	11	12	15.6
5		Real/real	2.5	1.17	268	4426	12842	21	60	11	63	39.3
6	1BKR	Real/ideal	2.5	1.1	262	5841	12184	0	25	11	3	5.1
7		Real/real	2.5	1.1	262	5749	12184	17	26	11	6	7.6
8		Real/real	2.5	1.1	262	5598	12184	21	51	12	56	35.8
9	1TMY	Ideal/ideal	3.0	2.0	262	2426	7737	0	61	13	37	31.9
10		Ideal/ideal	3.0	2.0/1.5	262	3653	9143	0	24	12	9	11.2
11		Real/ideal	2.5	1.9	262	2704	8960	0	78	12	60	37.8

In a new ideal/ideal-type calculation, we have kept all initial data (phased E at 3.0 Å and observed moduli at 2.0 Å) identical to these of no. 9 but, unlike all previous calculations, we have used from the very beginning a Ψ set extended beyond the resolution of the observed E set, down to 1.5 Å resolution. Then we have allowed the additional calculation by equation (1) of E also down to 1.5 Å resolution. These additional E (with calculated moduli and phases) are in turn introduced into all subsequent calculations, provided, of course, that they pass successfully the same tests as the observed E (AMIN). The electron-density maps relative to the residues 99–101 for no. 10 confirm that many atoms are now almost resolved (Fig. 4).

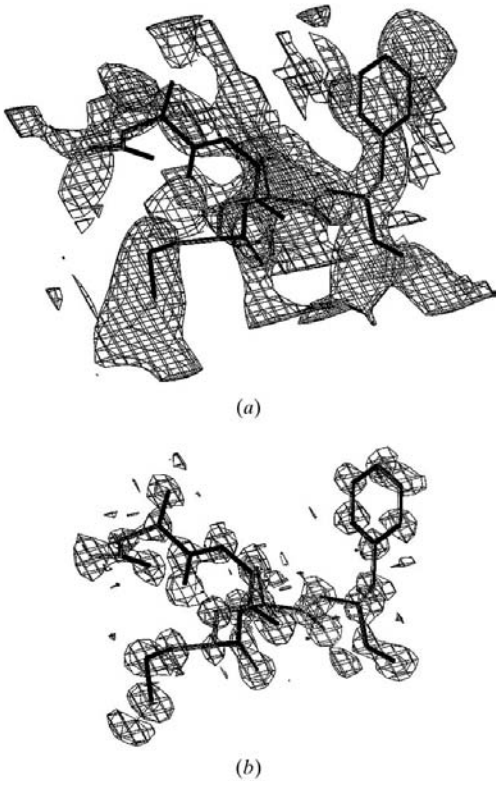

Figure 3
Electron-density maps of residues A33–A37 of protein 1BKR. Maps are contoured at 2.0σ. (a) Map plotted by the initial phased set at resolution 2.5 Å. (b) Map plotted by the extended phased set at resolution 1.1 Å (Table 2, no. 6).

A striking result is that not only are these 'super-resolution' E now correctly phased (MPE of 25°), but all other E are shifted as well towards the correct values, as shown in Table 3. This emphasizes the paramount importance of the choice of the auxiliary variables that can be set independently of the resolution of the observed E: choosing Ψ in a resolution range higher than that of the observed E is already a significant preparative step towards phasing of E in a super-resolution range.

6. The crystallographic symmetry test

An important feature of the *TWIN* algorithm is the decoupling between the E, bearing the observed $E_{\mathbf{H}}$ moduli information, and the auxiliary variables Ψ which alone control the phasing procedure. This feature enabled us to develop a new FOM for each reflection and a new overall evaluation test based on crystallographic symmetry. The test consists of the symmetry mean phase error, S_MPE, and the symmetry mean modulus error, S_R_{mod}, which are a measure of the inconsistency between the calculated phases and magnitudes by equation (1) for symmetry-related reflections. In space group $P2_1$, we have

$$\mathrm{S\_MPE} = \sum_{\mathbf{H}} |\varphi_{\mathbf{H}} - \varphi_{\mathbf{RH}} \pm k\pi|, \qquad (7a)$$

$$\mathrm{S\_R_{mod}} = \sum_{\mathbf{H}} \big[|E_{\mathbf{RH}}| - |E_{\mathbf{H}}|\big] \Big/ \sum_{\mathbf{H}} |E_{\mathbf{H}}|, \qquad (7b)$$

where $\varphi_{\mathbf{RH}} = \varphi_{\mathbf{H}} - k\pi$, k is the Miller index of the reflection, $\mathbf{H} = (hkl)$ and $\mathbf{RH}$ is the symmetry-equivalent reflection 2_1.

The Ψ are not restricted by theory to obey the symmetry constraints and, therefore, the E calculated by equation (1) in step (b) are not symmetry restricted either. Thus, the decrease of S_MPE and S_R$_{\rm mod}$ throughout the iterations is likely to reflect the correctness of the phasing procedure at each iteration.

A particularly striking situation occurs at the end of the preliminary step (*a*), before the first iteration of step (*b*): this could be termed as an extreme dilution of the symmetry information contained in the initial known set. This dilution results in a very weak degree of fulfilment of the space-group-symmetry relations for the Ψ set. This can be quantitatively evaluated by computing the S_MPE pertaining to the Ψ set. For instance, in set no. 1 of Table 2, the S_MPE before step (*b*) for the 9257 Ψ is 87°, that is indeed very close to the random value of 90°.

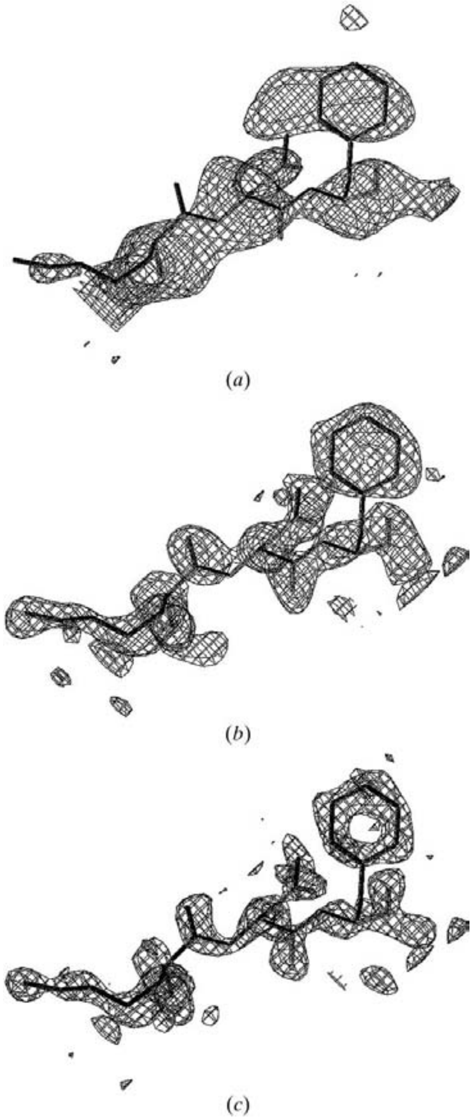

Figure 4
Electron density maps of residues 99–101 of protein 1TMY. Maps are contoured at 2.3σ. (*a*) Map plotted by the initial phased set at resolution 3.0 Å. (*b*) Map plotted by the extended phased set at resolution 2.0 Å. (*c*) Map plotted by the 'super-resolution' extended set at 1.5 Å resolution (Table 2, no. 10).

Table 3
Analysis of the final results on phase extension for protein 1TMY including super-resolution: number of phases determined by the present algorithm and corresponding MPE.

The initial known set consists of 262 phases at 3.0 Å resolution. 2118 moduli are used at 2.0 Å resolution. Super-resolution concerns the 2.0 < *d* < 1.5 Å shell where no observed moduli are used. 9143 Ψ at 1.5 Å resolution have been used throughout the calculations, of which 4932 are located outside the observed sphere.

Resolution (Å)	No. of extended *E*'s	MPE of extended set (°)
$d > 3.0$	410	15.0
$3.0 \le d < 2.0$	1154	21.5
$2.0 \le d < 1.5$	2089	25.6

The algorithm subsequently generates *E* which are progressively endowed with values closer to the correct symmetry relations for both moduli and phases. Therefore, both S_MPE and S_R$_{\rm mod}$ can now be used as a FOM for acceptance for each reflection issued throughout the iterations in step (*b*), as well as a test for the success of the whole algorithm. The last columns of Table 2 show the final values of S_MPE and of S_R$_{\rm mod}$. It appears that their values provide an indication for the correctness of the phase-determination process. Note that these are preliminary results and it can be expected that further work will improve the reliability of the test. From the present results, we can, however, consider that an S_MPE of 40° is a tentative cut-off value for acceptance of a solution in a multisolution procedure. Fig. 5 provides a detailed description of the variation of several 'symmetry indicators' during the phase extension process. The information provided by S_MPE can be extended by that from Ψ_S_MPE,

$$\Psi\_{\rm S\_MPE} = \sum_{\mathbf{K}} \left| \omega_{\mathbf{K}} - \omega_{\mathbf{RK}} \pm k\pi \right|, \tag{8}$$

where $\omega_{\mathbf{K}}$ is the phase of $\Psi_{\mathbf{K}}$. The curve is smoother on account of the very large number of contributors in (8) from the very beginning, as compared with that of (7*a*).

As a final remark, it could be considered that the symmetry information would be more useful if it were fed into the algorithm from the very beginning and, subsequently, strictly obeyed for *E* throughout the procedure, *i.e.* by setting and keeping Ψ so as to always obey the symmetry relations.

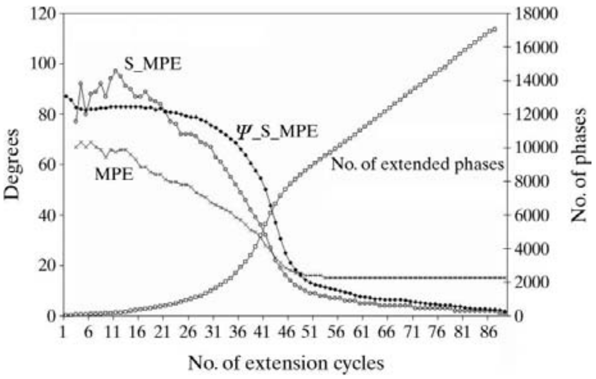

Figure 5
Variation of several 'symmetry indicators' during the phase extension process for set no. 1 of Table 2.

Table 4
Selection of the best refined phase sets for 150 random Ψ sets for a 41 non-H atom structure (notations and data for the structure are given in H–T).

	Minimization functions			
	M_{mod} $M_{mixed\ triplets}$		M_{mod} $M_{mixed\ triplets}$ M_{quart}	
Random set no.	No. of E's	MPE (°)	No. of E's	MPE (°)
1	126	19	173	31
	187	35	261	34
			224	35
37	151	27	193	26
	178	30	210	27
			266	36

However, the symmetry tests appear to be very useful for the evaluation of the correctness of the phasing process (especially for a multisolution algorithm), and this new criterion in direct methods probably produces a hypothetical extra power provided by constraining the symmetry of Ψ. It is important to emphasize that the classical R factor is not a good indicator, as shown in §3. The sigmoid variation of Ψ_S_MPE confirms the cut-off value of approximately 40° deduced from Table 2.

7. Towards *ab initio* phase determination: multisolution algorithms

First we have examined the effect of the progressive decrease of the number of initially phased E for protein structures and have found that 200–270 initially phased E seem to represent the minimal initial information necessary for phase extension by the present algorithm for structures containing close to 1000 atoms. Further improvements are necessary for a reduction of the ratio 'number of known phases/number of atoms' beyond the above range of about 1/3 to 1/4. Such improvements could be provided by the use of a multisolution procedure.

7.1. *Ab initio* calculations for small structures

The multisolution approach in modern direct methods has been proved very successful in solving crystal structures. Thus, we tested the multiple generation of random starting sets of Ψ auxiliary variables and their subsequent treatment by the *TWIN* algorithm. For these test calculations, we have used ideal E for a structure with 41 atoms in $P1$ (Psicharis, Mentzafos & Terzis, unpublished data; see also H–T). The experiments have shown that 150 random starts were able to produce several acceptable solutions with an unweighted MPE in the range 25–35° for 150–200 E. The results are summarized in Table 4 showing the lowest MPEs and corresponding numbers of E for the best sets out of 150 random trials. It should be noted that the inclusion of the M_{quart} function (see H–T) into the minimization procedure further improves the results. Note that for the other starting sets the MPE is considerably higher and the number of phased E is smaller than the corresponding values of the above table.

This research was supported by NATO–Brussels International Scientific Exchange Programmes linkage grants LG 951447.

References

Banuelos, S., Saraste, M. & Djinovic Carugo, K. (1998). *Structure*, **6**, 1419.

Barret, A. N. & Zwick, M. (1971). *Acta Cryst.* A**27**, 6–11.

Bethanis, K., Tzamalis, P., Hountas, A., Mishnev, A. & Tsoucaris, G. (1997). Poster communication. 25th NATO ASI Summer School, Erice, Sicily, Italy, 22 May–2 June 1997.

Bezborodova, S. I., Ermekbaeva, L. A., Shlyapnikov, S. V., Polyakov, K. M. & Bezborodov, A. M. (1988). *Biochem. USSR*, **53**(6–2), 837–845.

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P. & Wu, C.-W. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 4175–4179.

Collins, D. M. (1982). *Nature (London)*, **298**, 49–51.

Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1988). *Acta Cryst.* A**44**, 353–357.

Hountas, A. & Tsoucaris, G. (1995). *Acta Cryst.* A**51**, 754–763.

Jones, T. A. & Kjeldgaard, M. (1995). *O. Molecular Graphics Program*, Version 5.10. Uppsala University, Sweden.

Podjarny, A. D., Bhat, T. N. & Zwick, M. (1987). *Annu. Rev. Biophys. Chem.* **16**, 351–373.

Rango, C. de, Mauguen, Y., Tsoucaris, G., Dodson, E. J., Dodson, G. G. & Taylor, D. J. (1985). *Acta Cryst.* A**41**, 3–17.

Roversi, P., Irwin, J. J. & Bricogne, G. (1998). *Acta Cryst.* A**54**, 971–996.

Sayre, D. (1974). *Acta Cryst.* A**30**, 180–184.

Shaefer, M., Schneider, T. R. & Sheldrick, G. (1998). *Structure*, **4**, 1509–1515.

Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1996). *Acta Cryst.* A**52**, C64–C65.

Tsoucaris, G., Mishnev, A. & Hountas, A. (1996). *Acta Cryst.* A**52**, C54.

Usher, K. C., De La Cruz, A., Dahlquist, F. N., Swanson, R. V., Simon, M. I. & Remington, S. J. (1997). Deposited with Protein Data Bank, PDB ID=1TMY.

Woolfson, M. M. & Yao, J.-X. (1988). *Acta Cryst.* A**44**, 410–413.

Zhang, K. Y. & Main, P. (1990). *Acta Cryst.* A**46**, 41–46.