Acta Cryst. (1997). A**53**, 789–798

Improving Direct-Methods Phases by Heavy-Atom Information and Solvent Flattening

C. GIACOVAZZO[a]* AND D. SILIQI[a,b]

[a]*Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and* [b]*Department of Inorganic Chemistry, Tirana University, Tirana, Albania. E-mail: criscg01@area.ba.cnr.it*

(*Received 23 January 1997; accepted 30 June 1997*)

Abstract

In a recent series of papers, a procedure has been described for phasing all the reflections up to derivative resolution by using triplet phase relationships [Giacovazzo, Cascarano & Zheng (1988). *Acta Cryst.* A**44**, 45–51]. The resulting electron-density maps show good correlation with the correct maps but usually are not straightforwardly interpretable. In this paper, the quality of the phases is improved by: (*a*) exploiting the information on the heavy-atom structure, which becomes available as soon as protein phases are available; (*b*) applying a suitable solvent-flattening procedure (*FLEX*), which proves highly effective in reducing the phase error.

1. Symbols and abbreviations

Symbols and notations are basically the same as in papers I–VI of a series (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995; Giacovazzo & Gonzalez-Platas, 1995; Giacovazzo, Siliqi & Gonzalez-Platas, 1995; Giacovazzo, Siliqi, Gonzalez-Platas, Hecht, Zanotti & York, 1996).

2. Introduction

Traditional direct methods (*e.g.* Sayre's equation, tangent formula, determinantal approaches *etc.*) cannot solve *ab initio* crystal structures of common size proteins. Supplementary information is needed which may be provided, for example, by derivative data. The integration of direct methods with isomorphous replacement techniques (SIR case) was first accomplished by Hauptman (1982), who derived the joint probability distribution of the six structure factors

$$P(E_{\mathbf{h}}, E_{\mathbf{k}}, E_{\mathbf{h}-\mathbf{k}}, G_{\mathbf{h}}, G_{\mathbf{k}}, G_{\mathbf{h}-\mathbf{k}}),$$

E and G being the structure factors of the native protein and of the derivative, respectively. The conditional probability formula estimating the triplet phase invariant (constituted by native phases)

$$\Phi = \phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h}-\mathbf{k}}$$

given the six moduli $R_i = |E_i|$ and $S_i = |G_i|$ is of the von Mises type. Its concentration parameter, stating the reliability of the phase indication, is a complicated expression of the six moduli. The Hauptman mathematical procedure was reconsidered by Giacovazzo, Cascarano & Zheng (1988), who obtained a simpler distribution:

$$P(\Phi) \cong [2\pi I_o(A)]^{-1} \exp(A \cos \Phi), \qquad (1)$$

where

$$A = 2[\sigma_3/\sigma_2^{3/2}]_p R_{\mathbf{h}} R_{\mathbf{k}} R_{\mathbf{h}-\mathbf{k}} + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_{\mathbf{h}} \Delta_{\mathbf{k}} \Delta_{\mathbf{h}-\mathbf{k}}$$

and $\Delta = (|F_d| - |F_p|)/\Sigma_H^{1/2}$ is the pseudo-normalized difference (with respect to the heavy-atom structure).

Papers I–VI were devoted to describing a procedure for phasing all the reflections up to derivative resolution based on the application of relation (1). Applications to experimental data were successful, so that the procedure can be considered to be competitive with traditional SIR techniques. It may be briefly described as follows.

(*a*) *Normalization step* (see papers III and V). The usual Wilson plot technique is used to scale native data. Then a differential Wilson plot (Blundell & Johnson, 1976) is applied for approximately scaling the derivative on native data. The final scaling factor is found by using the properties of the $P(\Delta)$ distribution, which proves to be basic mostly when the derivative resolution is 4 Å or less.

(*b*) *Phasing step* (see papers III and VI). The set of reflections up to derivative resolution is divided in batches. For the first batch (usually constituted by 800–1000 reflections with the largest values of $|\Delta|$), a starting set of phases is generated by a random process (Baggio, Woolfson, Declercq & Germain, 1978) to which a weighted tangent formula, arising from equation (1), is applied. Among the various trials produced by the multisolution approach, that with the highest value of CFOM [CFOM is the combined figure of merit; see step (*c*)] is chosen. Then batches of about 200 reflections, chosen in decreasing order of $|\Delta|$, are progressively phased *via* a phase-extension procedure from batch no 1.

(*c*) *Identification of the correct solution.* Figures of merit are used to identify the correct solution. If the derivative is of sufficiently high quality, there is no difficulty in finding the correct solution among the various trials. A supplementary check requires the use of

the difference Fourier synthesis (for the first batch of reflections only) in order to discard trials with high figures of merit but locating heavy atoms on allowed origins (see paper V).

The above procedure provides electron-density maps that are at least as informative as those generated by application of traditional SIR techniques (see paper VI). The results may be described as: (i) all the reflections up to derivative resolution can be phased. To each phase, a weight is associated that is related to the reliability of the phase indication, and arises from the application of the tangent formula; (ii) the maps are directly interpretable only when a high-quality derivative is available; (iii) bad isomorphism and heavy errors in measurements can hinder the success of the procedure. However, even in this case, an interesting correlation between the trial map and the 'true' map can be found; (iv) The procedure does not require the previous knowledge of the heavy-atom structure. Thus, protein phases can be obtained in the absence of such information, which is instead basic for traditional SIR techniques. This last property is of concern in this paper.

If phases are obtained without knowing the heavy-atom structure, can its prior knowledge improve them? This question is of non-negligible interest: indeed, the refinement of the heavy-atom structure is a central tool for phase refinement in traditional SIR techniques (Cura, Krishnaswamy & Podjarny, 1992; Rould, Perona & Steitz, 1992). Part of this paper will be devoted to our attempts to improve phases obtained *via* equation (1) by using the prior information on the heavy-atom structure. The second part of the paper will be devoted to a solvent-flattening procedure, which shows interesting features of effectiveness and automatism. It starts once the procedure described in papers I–VI stops, and aims at extending and refining phases up to derivative resolution.

3. About the use of the heavy-atom structure

It has been emphasized that our direct-methods procedure works in the absence of any information on the heavy-atom structure. A decision however is required by the user when the procedure is started. In the normalization step, the application of both the differential Wilson plot and the $P(\Delta)$ distribution depend on the ratio $[\sigma_2]_H/[\sigma_2]_p$. While $[\sigma_2]_p$ is approximately known, no information on $[\sigma_2]_H$ is available. The ratio will depend on the number of heavy-atom sites per asymmetric unit and on the site occupancy. In the applications described here, we will use the same test structures and the corresponding experimental data employed in papers II–IV. The main characteristics of these data are shown in Table 1. In the applications described in papers II–IV, we assumed one heavy-atom site per asymmetric unit for APP, M-FABP, E2 and NOX and two sites for BPO, as in the published heavy-atom models. In contrast to this, we always fixed to unity the occupancy of such sites. Our

Table 1. *Relevant parameters for diffraction data of the test structures*

Structure code	Native RES (Å)	Native NREFL	Heavy atom	Derivative RES (Å)	Derivative NREFL
APP[a]	0.99	17058	Hg	2.00	2108
BPO[b]	2.35	23956	Au	2.78	15741
E2[c]	2.65	10391	Hg	3.00	9581
M-FABP[d]	2.14	7595	Hg	3.00	7125
NOX[e]	3.00	4619	Pt	3.00	4520

References: (*a*) Glover *et al.* (1983); (*b*) Hecht, Sobek, Haag, Pfeifer & Van Pee (1994); (*c*) Mattevi *et al.* (1992); (*d*) Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992); (*e*) Hecht, Erdmann, Park, Sprinzl & Schmid (1995).

Table 2. *Mean phase error (ERR) for the test structures up to derivative resolution*

$[\sigma_2]_H/[\sigma_2]_p$ parameters correspond to unitary occupancy factors of the heavy-atom sites. NREFD is the number of phased reflections up to derivative resolution. CORR is the correlation factor between direct-methods map (derivative resolution) and 'true' map (native resolution).

Structure code	$[\sigma_2]_H/[\sigma_2]_p$	NREFD	ERR (weighted)	CORR
APP	0.456	1850	61 (57)	0.3927
BPO	0.062	12774	57 (52)	0.4490
E2	0.078	6575	57 (52)	0.5121
M-FABP	0.128	5456	64 (61)	0.3733
NOX	0.081	4066	73 (69)	0.3129

direct-method procedure yielded the results in Table 2. NREFD is the number of phased reflections up to derivative resolution, ERR is the phase error of the assigned phase values (with the weighted phase error in parentheses), CORR is the correlation of our electron-density map ρ (calculated by directly phased reflections up to derivative resolution) with the 'true' map $\rho_{\rm mod}$ corresponding to the published phases (all reflections up to native resolution included). CORR has been calculated according to:

$$\mathrm{CORR} = \frac{\langle \rho \rho_{\rm mod} \rangle - \langle \rho \rangle \langle \rho_{\rm mod} \rangle}{(\langle \rho^2 \rangle - \langle \rho \rangle^2)^{1/2} (\langle \rho_{\rm mod}^2 \rangle - \langle \rho_{\rm mod} \rangle^2)^{1/2}}.$$

The arbitrary choice of the $[\sigma_2]_H/[\sigma_2]_p$ parameters may be corrected as follows. Once phases are available, a difference Fourier synthesis with coefficients $(F_d - F_p)\exp(i\phi_p)$ is calculated, which provides approximate heavy-atom structure parameters: these are refined according to Dickerson, Kendrew & Strandberg (1961). The scaling of derivative on native data is performed by exploiting the $P(\Delta)$ curves corresponding to the new $[\sigma_2]_H/[\sigma_2]_p$ values. New Δ's arise, which are then used in the phasing process. This time we do not need to enter into a multisolution procedure: it suffices to refine by a tangent process the phases obtained *via* the old Δ's. The results are in Table 3. Modest improvement of ERR and/or CORR values is obtained for APP and M-FABP. For

these structures, the new $[\sigma_2]_H/[\sigma_2]_p$ values are very different from the old ones: consequently, the corresponding $P(\Delta)$ curves are also quite different (see Fig. III.2). For BPO and E2, the refined values of $[\sigma]_H/[\sigma]_p$ do not substantially modify the previous situation, while for NOX it becomes slightly worse.

It may be concluded that the prior knowledge of the parameter $[\sigma_2]_H/[\sigma_2]_p$ is not critical for the success of the phasing procedure, which can safely work by using a rough approximation of its true value. However, when the 'true' value of $[\sigma_2]_H/[\sigma_2]_p$ is known, this may remarkably improve the efficiency of the procedure, provided the initial value of $[\sigma_2]_H/[\sigma_2]_p$ was very wrong.

A new perspective can be considered: the distribution (1) has been derived in the absence of any information on the heavy-atom structure. Can this supplementary information be used for improving the probability distribution of the triplet phases? An affirmative answer should make the tangent procedure more effective and lead to a remarkable reduction of the average phase error. Proposals for incorporating the heavy-atom structure information into the triplet phase distributions were made by various authors. The most significant are those suggested by Fortier, Moore & Fraser (1985) and by Klop, Krabbendam & Kroon (1987). Both are based on the following background.

Once the heavy atoms have been located, the cosine moduli of the doublet invariants $\delta_i = \psi_i - \phi_i$, $i = 1, 2, 3$, can be estimated *via* the Carnot relation

$$\cos\delta = (|F_d|^2 - |F_p|^2 - |F_H|^2)/2|F_dF_p|. \qquad (2)$$

Accordingly, the distribution

$$P(\phi_1, \phi_2, \phi_3, \delta_1, \delta_2, \delta_3|\{R'_i, S'_i, i = 1, 2, 3\}) \qquad (3)$$

can be derived by a simple change of variable from the distribution

$$P(\phi_1, \phi_2, \phi_3, \psi_1, \psi_2, \psi_3|\{R'_i, S'_i, i = 1, 2, 3\}).$$

We have

$$\begin{aligned}
&P(\phi_1, \phi_2, \phi_3, \delta_1, \delta_2, \delta_3|\{R'_i, S'_i, i = 1, 2, 3\}) \\
&\cong (1/L)\exp\left\{\sum_{i=1}^{3} 2R'_iS'_i\cos\delta_i + 2[\sigma_3/\sigma_2^{3/2}]_p\right. \\
&\times R'_1R'_2R'_3\cos\Phi + 2[\sigma_3/\sigma_2^{3/2}]_H[-R'_1R'_2R'_3\cos\Phi \\
&+ S'_1R'_2R'_3\cos(\Phi + \delta_1) - R'_1S'_2R'_3\cos(\Phi + \delta_2) \\
&+ R'_1R'_2S'_3\cos(\Phi + \delta_3) + R'_1S'_2S'_3\cos(\Phi + \delta_2 + \delta_3) \\
&- S'_1R'_2S'_3\cos(\Phi + \delta_1 + \delta_3) \\
&- S'_1S'_2R'_3\cos(\Phi + \delta_1 + \delta_2) \\
&\left.+ S'_1S'_2S'_3\cos(\Phi + \delta_1 + \delta_2 + \delta_3)]\right\}. \qquad (4)
\end{aligned}$$

In Fortier *et al.*'s (1985) method, the signs of the δ_i's were supposed unknown, then from equation (4) the conditional distribution

$$P(\Phi|\{R'_i, S'_i, |\delta_i|, i = 1, 2, 3\}) \qquad (5)$$

may be obtained as a weighted sum of the eight distributions $P(\Phi|\{R'_i, S'_i, |\delta_i|, i = 1, 2, 3\})$ corresponding to the eight sign combinations of the doublet invariants δ_1, δ_2 and δ_3. Applications of (5) to experimental data did not improve the results in Table 3: the reader is referred to a recent paper (Giacovazzo, Siliqi, Cascarano, Caliandro & Melidoro, 1997) for details.

Table 3. *Mean phase error (ERR) when the information on the occupancy of the heavy-atom sites has been exploited (data up to derivative resolution)*

$[\sigma_2]_H/[\sigma_2]_p$ parameters correspond to refined occupancy factors of the heavy-atom sites. NREFD is the number of phased reflections up to derivative resolution. CORR is the correlation factor between direct-methods map (derivative resolution) and 'true' map (native resolution).

Structure code	$[\sigma_2]_H/[\sigma_2]_p$	NREFD	ERR (weighted)	CORR
APP	0.055	1863	59 (55)	0.4565
BPO	0.028	12673	57 (52)	0.4452
E2	0.021	6556	56 (52)	0.4968
M-FABP	0.015	5630	64 (60)	0.4069
NOX	0.041	3858	73 (69)	0.3020

Klop, Krabbendam & Kroon's (1987) proposal tries to exploit also the δ_i signs: indeed, estimates of them are available when ϕ_p and the heavy-atom structure are known. The conditional distribution

$$P(\phi_1, \phi_2, \phi_3, \delta_1, \delta_2, \delta_3|\{R'_i, S'_i, i = 1, 2, 3\})$$

may then be used, from which

$$P(\Phi|\{R'_i, S'_i, \delta_i, i = 1, 2, 3\}) \cong (1/L)\exp[A_n\cos(\Phi - \xi_n)] \qquad (6)$$

is obtained, where

$$\begin{aligned}
A_n\cos\xi_n &= 2[\sigma_3/\sigma_2^{3/2}]_pR'_1R'_2R'_3 \\
&+ 2[\sigma_3/\sigma_2^{3/2}]_H[-R'_1R'_2R'_3 + S'_1R'_2R'_3\cos\delta_1 \\
&+ R'_1S'_2R'_3\cos\delta_2 + R'_1R'_2S'_3\cos\delta_3 \\
&- R'_1S'_2S'_3\cos(\delta_2 + \delta_3) - S'_1R'_2S'_3\cos(\delta_1 + \delta_3) \\
&- S'_1S'_2R'_3\cos(\delta_1 + \delta_2) \\
&+ S'_1S'_2S'_3\cos(\delta_1 + \delta_2 + \delta_3)] \\
A_n\sin\xi_n &= 2[\sigma_3/\sigma_2^{3/2}]_H[S'_1R'_2R'_3\sin\delta_1 + R'_1S'_2R'_3\sin\delta_2 \\
&+ R'_1R'_2S'_3\sin\delta_3 - R'_1S'_2S'_3\sin(\delta_2 + \delta_3) \\
&- S'_1R'_2S'_3\sin(\delta_1 + \delta_3) - S'_1S'_2R'_3\sin(\delta_1 + \delta_2) \\
&+ S'_1S'_2S'_3\sin(\delta_1 + \delta_2 + \delta_3)].
\end{aligned}$$

ξ_n is the most probable value of Φ: if the δ_j's are very close to zero, then $A_n \equiv |A|$ and $\xi_n = 0, \pi$. This occurs when $|F_H|$ is negligible with respect to $|F_p|$ and $|F_d|$, and occasionally when F_H and F_p have equal or opposite

phase value. In order to have more insight into the properties and limits of equation (6), we note:

(*a*) Equation (6) is expected to be more informative than equation (1) because it is able to take into account the prior information on the signs of δ_1, δ_2 and δ_3. However, in our procedure such signs can only be calculated if ϕ_p and the heavy-atom structure are known. Therefore, the efficiency of equation (6) in practice will depend on the accuracy with which the phases ϕ_p and ϕ_H are known. To a first approximation, ϕ_H may be considered known with good accuracy: accordingly, ξ_n will depend on the assigned value of ϕ_p and the application of (6) will tend to generate the same errors as the application of (1).

(*b*) The distribution (1) is obtained in the absence of any information on F_H. Incorporating such information into (1) *a posteriori* [that is, after the mathematical form of (1) has been fixed on assuming that E_H is unknown] is a practical but not fully correct way for improving the effectiveness of the distribution. On the other hand, there is no sense in trying to derive the distribution

$$P(E_{p_1}, E_{p_2}, E_{p_3}, E_{d_1}, E_{d_2}, E_{d_3}, E_{H_1}, E_{H_2}, E_{H_3}) \qquad (7)$$

in order to calculate (6) as a conditional distribution of (7). Indeed, E_{p_i}, E_{d_i} and E_{H_i} are algebraically related by

$$E_{d_i} = E_{p_i} + E_{H_i} \qquad (8)$$

and therefore (6) would be Dirac-delta-function-like, assuming zero values when (8) is not verified and infinite when it is fulfilled.

We applied (6) both to calculated and to experimental data. According to our statements in (*a*), in the first case (6) works much better than (1) since ϕ_p is exactly known. On the contrary, when (6) is applied to experimental data, it is not able to improve the results in Table 3.

There are three supplementary ways to profit by the prior knowledge of the heavy-atom structure: (i) to recognize cross-over cases; (ii) to use the information on Φ_H; (iii) to employ better estimates of Δ. Case (i) is of marginal practical interest because very few cross-over cases were recognized in our test structures. How this case may be managed by (6) is described in Appendix *A*.

Case (ii) was considered in paper VI. It was shown that the assumption $\Phi_H \simeq 0$ is implicit in (1), and that the number of cases in which such a condition is violated is not negligible. Equation (1) could then be modified into

$$P(\Phi) \cong [2\pi I_o(A)]^{-1} \exp[A \cos(\Phi - \Phi_H)], \qquad (9)$$

so making more explicit the correlation between SIR and direct-methods techniques. While equation (1) suggests that the expected value of Φ_H is 0 if $A > 0$, π if $A < 0$, equation (9) indicates that the expected value is Φ_H if $A > 0$, $(\Phi_H + \pi)$ if $A < 0$. In order to check the relative reliability of (1) and (9), we estimated the triplet phases found among the 800 reflections of M-FABP with the largest $|\Delta|$ values. In Table 4, NTRP and NTRN are the number of triplets with $A > 0$ and $A < 0$, respectively, NCOSP and NCOSN are the numbers of triplets with $\cos \Phi_H \geq 0$ and $\cos \Phi_H < 0$, respectively, and $\langle|\Phi|\rangle_{\text{NCOSP}}$ and $\langle|\Phi|\rangle_{\text{NCOSN}}$ the average absolute phase values for the NCOSP and NCOSN triplets, respectively. Table 4 clearly shows that triplets with $\cos \Phi_H < 0$ are better estimated *via* (9) than *via* (1). Similar results, not shown for brevity, are obtained for all the test structures.

Table 4. *Triplet reliability according to equations (1) and (9)*

See the main text for the symbols.

M-FABP

NTRP	NCOSP	$\langle\|\Phi\|\rangle_{\text{NCOSP}}$	NCOSN	$\langle\|\Phi\|\rangle_{\text{NCOSN}}$
15546	13545	68°	2001	93°
NTRN	NCOSP	$\langle\|\Phi\|\rangle_{\text{NCOSP}}$	NCOSN	$\langle\|\Phi\|\rangle_{\text{NCOSN}}$
13918	12249	108°	1669	86°

BPO

NTRP	NCOSP	$\langle\|\Phi\|\rangle_{\text{NCOSP}}$	NCOSN	$\langle\|\Phi\|\rangle_{\text{NCOSN}}$
25290	20180	60°	5110	108°
NTRN	NCOSP	$\langle\|\Phi\|\rangle_{\text{NCOSP}}$	NCOSN	$\langle\|\Phi\|\rangle_{\text{NCOSN}}$
24509	19455	119°	5054	71°

We devised two ways to apply (9). The first excludes from the phasing process the triplets with $\cos \Phi_H < 0$ [the rationale is that equations (1) and (9) disagree in this case]. The experimental results were disappointing: the average absolute phase error at the end of the phasing procedure remains practically unchanged. The second way actively applies (9) as it is. In Table 5, we show statistical calculations on the triplets found among the 800 reflections of M-FABP with the largest value of $|\Delta|$. Triplet phases are estimated *via* (1) and (9): NR is the number of triplets with $|A| >$ ARG, % is the percentage of triplets with correctly estimated $\cos \Phi$ value [those with $A \cos \Phi_{\text{est}} > 0$ when (1) is used, with $A \cos(\Phi_{\text{est}} - \Phi_H)$ if (9) is used].

Table 5 shows that the estimates provided by (9) are more accurate: however, the phasing process is unable to transform such a larger accuracy into better phases. The 800 M-FABP reflections are phased by (9) with an average error of 40.8° against 39.6° obtained *via* (1). A more favourable situation is that described in Table 5 for BPO. The larger accuracy of (9) positively influences the phasing process. The 1000 reflections of BPO (those with the largest $|\Delta|$ values) are phased with an absolute average error of 30.3° if (1) is used and of 28.9° if (9) is used. APP and E2 are in the intermediate situation: the average error for APP is 21.3° for both (1) and (9), while for E2 the error is 28.0° when (1) is used and 27.4° when

Table 5. *Triplet phase statistics according to equations (1) and (9)*

See the main text for the symbols.

M-FABP

		Equation (1)		Equation (9)	
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	%	$\langle\|\Phi\|\rangle$ (°)
0.8	29539	64	73	68	69
2.0	6155	69	67	72	64
3.8	113	68	65	68	69

BPO

		Equation (1)		Equation (9)	
ARG	NR	%	$\langle\|\Phi\|\rangle$ (°)	%	$\langle\|\Phi\|\rangle$ (°)
0.8	50000	67	70	82	52
1.6	2138	78	57	89	43
2.0	179	86	47	93	38

(9) is used. The above considerations suggests that the use of (9) instead of (1) can improve the phasing process: however, our tests show that improvement is marginal if any. For example, for BPO, when all the reflections up to derivative resolution are phased, the mean phase error is 56° (weighted value 52°), very close to that shown in Table 3.

We then tested technique (iii): to derive and apply better estimates of Δ by using the prior information on the heavy-atom structure. Once F_H and ϕ_H are available, from the right-hand side of (10) a better estimate of $|\Delta|$ may be derived:

$$|\Delta| = ||E_H^2|\cos(\phi_d - \phi_p) - 2|E_p'|\sin^2[(\phi_d - \phi_p)/2]|. \tag{10}$$

The supplementary condition

$$\text{if} \quad |\Delta| > |E_H|, \quad \text{then fix} \quad |\Delta| = |E_H| \tag{11}$$

was also applied. In our experimental tests, we applied (10) and (11) both together and separately. In all the cases, no valuable improvement has been found.

The conclusions that may be drawn from our numerous tests are the following: (*a*) The only valuable kind of supplementary information arising from the heavy-atom structure is the refined parameter $[\sigma_2]_H/[\sigma_2]_p$, which has been used for obtaining the phases analysed in Table 3. Such phases will constitute the starting point for the solvent-flattening procedure described in the next section. (*b*) No valuable supplemental information is obtained *via* the use of the so-called 'doublet invariants' δ_i, as noted by Klop *et al.* (1987) and Fortier *et al.* (1985).

4. The solvent-flattening procedure

Solvent-flattening methods (Wang, 1985; Leslie, 1987) are well known. They first determine a molecular envelope: then the electron density in the solvent is set to a constant value and the electron density in the protein region is constrained to be positive almost everywhere. The volume V_p of the protein region is usually estimated through the formula given by Mathews (1968). A cyclic procedure then starts according to which the modified electron-density map ρ is inverted and the resulting phases $\{\phi\}$ are combined with SIR or MIR phases. A new electron-density map is then recalculated and the cycle restarts.

We have devised a solvent-flattening procedure (from now on denoted *FLEX*) that may be automatically applied to phases provided by direct methods: it is quite effective for reducing the phase error and providing an improved electron-density map. Let us denote by NREFP the number of protein reflections up to the native resolution D. Then,

$$\text{NREM} = \text{NREFP} - \text{NREFD}$$

is the number of reflections that we have to phase through *FLEX*. The procedure consists of a variable number of *supercycles*: the first supercycle does not make use of the molecular envelope and is constituted by five *macrocycles*: the first of them is devoted to phase extension (up to native resolution) and to preliminary refinement, and is constituted by five *microcycles*. In each of them, the calculations $\rho \rightarrow \{\phi\} \rightarrow \rho$ are performed. Only a small fraction (say $V_f = 0.2V_p$) of the electron-density map is used in the step $\rho \rightarrow \{\phi\}$: such a fraction contains the pixels with the largest value of ρ. The phases obtained by Fourier inversion of the modified ρ map are combined with direct-methods (DM) phases *via* the tangent-like formula

$$\tan\phi_c = \left[\sum_{i=1,2} m_i w_i \sin\phi_i\right] \Big/ \left[\sum_{i=1,2} m_i w_i \cos\phi_i\right], \tag{12}$$

where ϕ_c is the combined phase value, $i = 1$ corresponds to DM phases, $i = 2$ to phases obtained *via* electron-density inversion, w_1 is the weight fixed by our DM procedure (see paper II), $w_2 = D_1(2|E_c'|)$, where E_c' is the value of E' obtained *via* electron-density inversion. In each microcycle, the $N_{\text{rem}}/5$ reflections having the largest value of $|E_c'|$ are phased.

Macrocycles from two to five are devoted to phase refinement. Consecutive macrocycles perform electron-density inversion by using increasing fractions V_f of the map:

$$V_f = 0.20, 0.25, 0.30, 0.35V_p.$$

Again, each macrocyle is constituted by five microcycles: in each of them, phases are combined according to (12) but now ϕ_1 corresponds to phases obtained at the end of macrocycle 1. Furthermore, a progressively larger weight for ϕ_2 is applied: in each first microcycle, $m_1 = 0.7$ and $m_2 = 1.3$, in subsequent microcycles, m_1 is diminished

by 0.05 and m_2 is increased by 0.05. Thus, in each fifth microcycle, $m_1 = 0.5$ and $m_2 = 1.5$.

In supercycle 2 and those following, the molecular envelope is calculated according to the Wang–Leslie method. The electron-density inversion is calculated by assigning unit weight to pixels in the protein region, weight equal to 0.50 to pixels in the solvent region.

In supercycle 2, the molecular envelope is calculated by using the average radius $R_1 = 8$ Å. 15 microcycles are performed: five by using $V_f = 0.20V_p$, five with $V_f = 0.35V_p$ and five with $V_f = 0.50V_p$. In each microcycle, the ratio (Rayment, 1983; Cannillo, Oberti & Ungaretti, 1983)

$$\mathrm{RAT} = ||F_o| - |F_c||/|F_o|$$

is calculated, where $|F_o|$ is the observed amplitude of the protein reflection. RAT is used to correct the weight w_2: if RAT > 0.5, $w_2 = D_1(2|E'_c|)$, otherwise (under the condition that $w_2 \leq 1.0$) $w_2 = 1.2D_1(2|E'_c|)$.

Further supercycles can be performed. For each ith supercycle ($i > 2$):

(*a*) The molecular envelope is calculated by using the average radius $R_i = (R_{i-1} - 2)$ Å. In the last supercycle, $R_{\mathrm{last}} = (D + 2)$ Å is used. For example, if $D = 1.5$ Å then $R_{\mathrm{last}} = 3.5$ Å.

(*b*) V_p is replaced by

$$V_{pi} = 0.9V_{pi-1} \quad (\text{with} \quad V_{p2} = V_p).$$

E.g., in supercycle 3, $R_3 = 6$ Å and $V_3 = 0.9V_p$. From supercycle 3 on, only ten microcycles per supercycle are performed, all with V_f constantly equal to $0.5V_{pi}$. Finally, the steps from supercycle 2 to the last supercycle are repeated by using only those macrocycles corresponding to $V_f = 0.5V_{pi}$.

The following considerations highlight the main features of *FLEX*:

(i) Quite recently, Abrahams & Leslie (1996) used an initial solvent-flattening mask which selected the estimated solvent fraction and 50% of the protein fraction. Such a generous solvent mask is recursively modified by removing pixels with values above the solvent cut-off, provided they are connected to the protein region. Our procedure starts by inverting a much smaller fraction of the electron-density map (corresponding to $V_f = 0.2V_p$). Only in the final stages of the procedure are larger fractions used (up to $0.50V_p$).

(ii) In standard density-modification procedures, $(2|F_o| - |F_c|)$ coefficient values are used (Read, 1986) for calculating the next map of the iterative modification procedure. We found F_o maps more useful: this is probably because in our procedure only a small fraction of the electron-density map is used for Fourier inversion.

(iii) The phase expansion does not proceed per resolution shell. The only criterion is the value of $|E'_c|$, which seemed slightly better than the Sim (1960) criterion (involving $2|E'_oE'_c|$).

(iv) In standard solvent-flattening procedures, phases $\{\phi_2\}$ arising from electron-density inversion are usually combined with SIR phases. In our procedure, DM phases are the target only for phases $\{\phi_2\}$ calculated in the first macrocycle of the first supercycle. In the rest of the procedure, the target phases $\{\phi_1\}$ are those obtained at the end of the first macrocycle of the first supercycle. The larger effectiveness of the new target is probably due to the fact that phases have smaller systematic errors than DM phases and better weighting factors.

(v) Weights m_1 and m_2 (for phase combination) progressively vary: as soon as refinement proceeds, the phases $\{\phi_2\}$ are considered more reliable than phases $\{\phi_1\}$.

(vi) The averaging radius R used for calculating the envelope varies from a maximum of 8 Å [typically used in the procedures derived from the one proposed by Wang (1985)] to a minimum value that depends on the native data resolution. In the case of APP, the minimum $R_{\mathrm{last}} \cong 3$ Å value was attained. The use of small averaging radii is very recent: a radius of 3.75 Å has been recently used by Abrahams & Leslie (1996) for the solution of the structure of bovine mitochondrial F_1 ATPase at 2.8 Å resolution. In our method, R decreases during the refinement process as soon as phases are expected to be sufficiently good. This improves the resolution of the envelope and creates small 'isles' of protein in the solvent region and larger 'lakes' of solvent in the protein region. This last feature prompted us to replace the protein volume V_p by V_{pi} in accordance with the progressive reduction of the averaging radius.

(vii) In our procedure, the solvent region is not set to a constant value (say ρ_{avg}) as in all traditional Wang-based flattening procedures. The weight 0.5 associated with pixels in the solvent region retards the flattening process (lower convergence to flatness) but allows the envelope to be improved if, in subsequent cycles, some prominent structural features of the solvent can become part of the protein region. Our procedure is an alternative to the 'flipping' technique proposed by Abrahams & Leslie, according to which, for all grid points within the solvent, the electron density is set to

$$\rho' = \rho_{\mathrm{avg}} + K_{\mathrm{flip}}(\rho - \rho_{\mathrm{avg}})$$

with $K_{\mathrm{flip}} = -1$.

(viii) Since DM phases (as well as SIR phases) are biased towards the heavy-atom structure, a spherical mask for reducing heavy-atom residuals in the calculated electron-density map is calculated.

5. Applications of *FLEX*

We applied *FLEX* to the five test structures: starting phase values and weights are those obtained after refinement of the occupancy of the heavy-atom sites (see Table 3). The procedure is applied in a default way,

Table 6. *Mean phase error (ERR) and the number of phased reflections (NREFP) up to native resolution after application of our solvent-flattening procedure*

CORR is the correlation factor between final map and 'true' map.

Structure code	Method	NREFP	ERR (weighted)	CORR
APP	*FLEX*	**17058**	**54 (48)**	**0.7780**
	Free-Sim	17040	77 (73)	0.5567
	Solomon	16360	85 (76)	0.4520
	Omit	15505	82 (77)	0.5015
BPO	*FLEX*	**23956**	**53 (46)**	**0.7391**
	Free-Sim	23949	57 (53)	0.6512
	Solomon	17859	55 (47)	0.6064
	Omit	23950	56 (50)	0.6694
E2	*FLEX*	**10391**	**41 (38)**	**0.8761**
	Free-Sim	10382	47 (42)	0.8101
	Solomon	10366	64 (52)	0.4888
	Omit	10383	51 (42)	0.7826
M-FABP	*FLEX*	**7589**	**54 (48)**	**0.6991**
	Free-Sim	7576	63 (59)	0.4898
	Solomon	7576	62 (57)	0.5144
	Omit	7589	61 (56)	0.5614
NOX	*FLEX*	**4619**	**72 (66)**	**0.4175**
	Free-Sim	4610	74 (70)	0.3380
	Solomon	4352	74 (70)	0.3495
	Omit	4615	75 (70)	0.3350

without any user intervention, and the results are summarized in Table 6.

APP: 1863 reflections (up to derivative resolution of 2.0 Å) were phased by direct methods with mean phase error ERR equal to 59° (weighted value 55°). The corresponding electron-density map shows a correlation with the 'true' refined map equal to 0.4565. DM phases are pseudo-centrosymmetrical (one symmetry-independent heavy-atom position in the space group *C*2). The solvent-flattening procedure breaks the pseudo-centrosymmetry and phases a total of 17 058 reflections, with ERR = 54° (weighted value 48°). The corresponding electron density is straightforwardly interpretable (CORR = 0.7780). In Fig. 1, the skeleton of APP is shown.

BPO: 12 673 reflections (up to derivative resolution of 2.8 Å) were phased by direct methods with mean phase error ERR equal to 57° (weighted value = 52°). The CORR value for the corresponding electron-density map is 0.4452. The *FLEX* procedure phases a total of 23 956 reflections, with an ERR value equal to 53° (weighted value = 46°). The corresponding electron-density map is straightforwardly interpretable (CORR = 0.7391). In Fig. 2, a section of the flattened map is compared with the corresponding section of the 'true' map.

E2: 6556 reflections (up to derivative resolution of 3.0 Å) were phased by direct methods with mean phase error ERR equal to 56° (weighted value 52°). The CORR value for the corresponding electron-density map is 0.4968. The *FLEX* procedure phases a total of 10 391 reflections, with an ERR value equal to 41° (weighted value 38°). The corresponding electron-density map is straightforwardly interpretable (CORR = 0.8761). In Fig. 3, a section of the flattened map is compared with the corresponding section of the 'true' map.

M-FABP: 5630 reflections (up to derivative resolution of 2.15 Å) were phased by direct methods with mean phase error ERR equal to 64° (weighted value 60°). The CORR value for the corresponding electron-density map is 0.4069. The *FLEX* procedure phases a total of 7589 reflections, with an ERR value equal to 54° (weighted value 48°). The corresponding electron-density map is straightforwardly interpretable (CORR = 0.6991). In Fig. 4, a section of the flattened map is compared with the corresponding section of the 'true' map.

NOX: 3858 reflections (up to derivative resolution of 3.0 Å) were phased by direct methods with mean phase error ERR equal to 73° (weighted value 69°). The CORR value for the corresponding electron-density map is 0.3020. The *FLEX* procedure phases a total of 4619 reflections, with an ERR value equal to 77° (weighted value 74°). The corresponding electron-density map is not immediately interpretable (DM phases are too noisy to constitute a sufficiently good starting point) and is marked by a CORR value equal to 0.4175.

In order to compare our method with other widely used flattening procedures, we processed our direct-

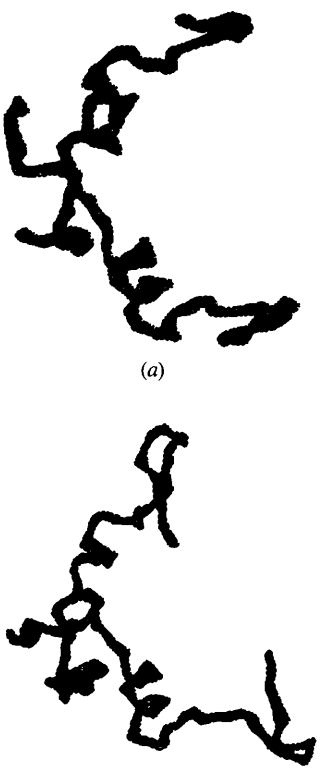
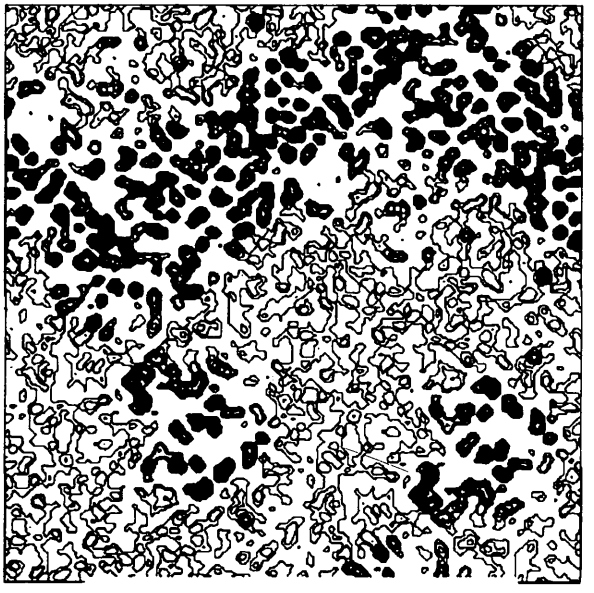


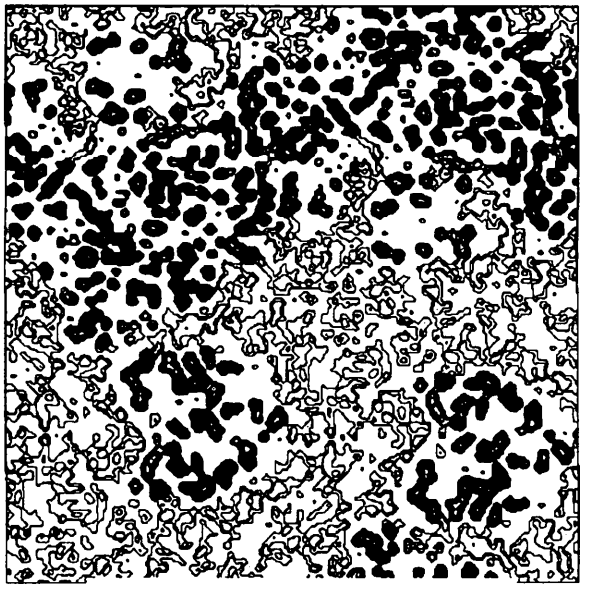
Fig. 1. APP. (*a*) Skeleton from the electron-density map obtained by applying *FLEX* to our direct-methods phases; (*b*) skeleton from the 'true' map (visualized by *RasMol* v2.3 by Roger Sayle).

methods phases by the program *dm*-1.6.9 of the CCP4-3.2 package (Collaborative Computing Project, Number 4, 1994) in the different modes: free-Sim mode, Solomon mode and reflection-omit mode (Leslie, 1987; Abrahams & Leslie, 1996; Cowtan & Main, 1996). The various modes contain both solvent-flattening and histogram-matching procedures and have been run according to the scripts shown in Table 7. For each test structure, the values of NREFP, ERR and CORR are given for a useful comparison. We observe that the efficiency of *FLEX* is always remarkably better.

It is worthwhile noting that we used all the programs in a 'default' way: therefore, we cannot claim that *dm* procedures cannot provide better results than those we

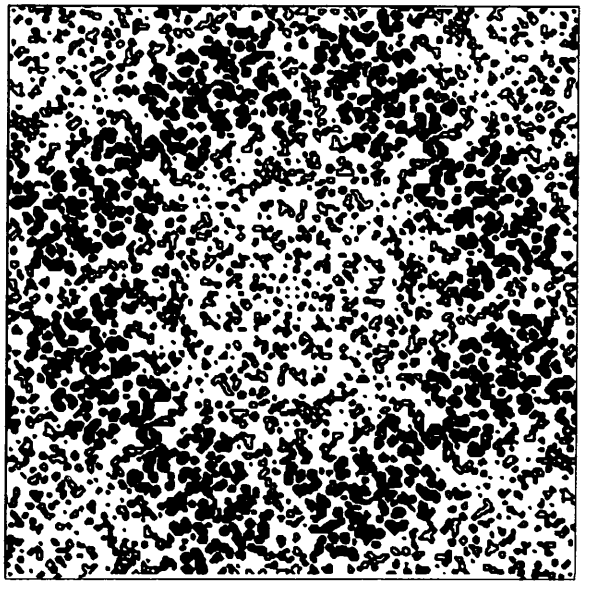


(*a*)

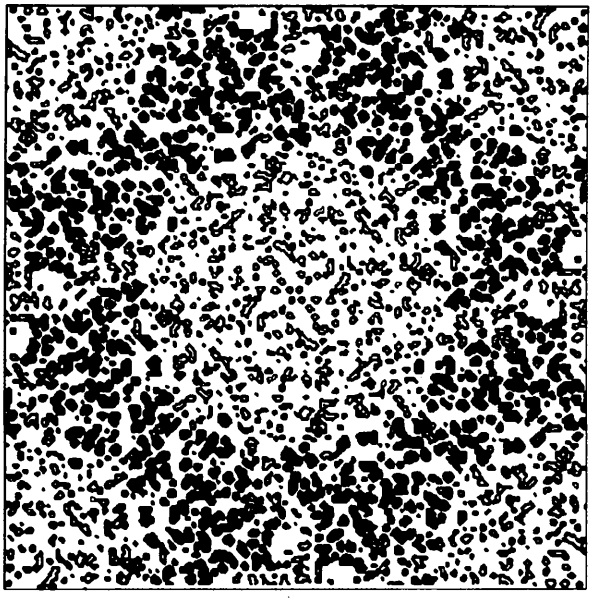


(*b*)

Fig. 2. BPO. (*a*) Section $y = 0$ of the electron-density map obtained by applying *FLEX* to our direct-methods phases; (*b*) section $y = 0$ of the true (obtained from the published model) map.



(*a*)



(*b*)

Fig. 3. E2. (*a*) Section $y = 0.3$ of the electron-density map obtained by applying *FLEX* to our direct-methods phases; (*b*) section $y = 0.3$ of the true (obtained from the published model) map.

Table 7. *Scripts used for dm program (to perform solvent flattening + histogram matching) in the three modes*

Free-Sim mode

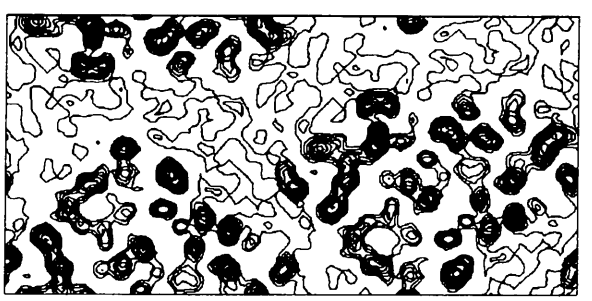
```
dm HKLIN struct_in.mtz HKLOUT struct_out.mtz << EOF
SOLC <solc>
MODE SOLV HIST
NCYC 200
LABIN FP = Fo SIGFP = SigFo PHIO = Ph_MD FOMO = W_MD
LABOUT PHIDM = Ph_Dm FOMDM = W_Dm
EOF
```

Solomon mode

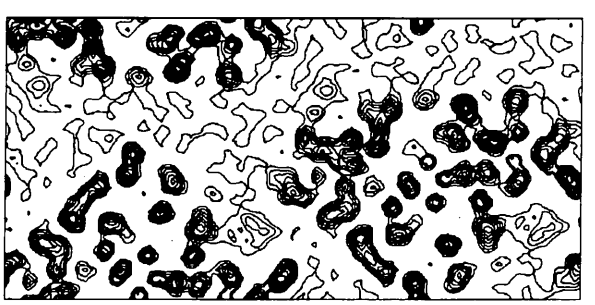
```
dm HKLIN struct_in.mtz HKLOUT struct_out.mtz << EOF
SOLC.. <solc>
MODE FLIP
NCYC AUTO
SCHEME ALL
COMBINE SIGMAA
LABIN FP = Fo SIGFP = SigFo PHIO = Ph_MD FOMO = W_MD
LABOUT PHIDM = Ph_Dm FOMDM = W_Dm
EOF
```

Reflection-omit mode

```
dm HKLIN struct_in.mtz HKLOUT struct_out.mtz << EOF
SOLC.. <solc>
MODE SOLV HIST
NCYC AUTO
SCHEME ALL
COMBINE OMIT
LABIN FP = Fo SIGFP = SigFo PHIO = Ph_MD FOMO = W_MD
LABOUT PHIDM = Ph_Dm FOMDM = W_Dm
EOF
```



(*a*)



(*b*)

Fig. 4. M-FABP. (*a*) Section $y = 0$ of the electron-density map obtained by applying *FLEX* to our direct-methods phases; (*b*) section $y = 0$ of the true (obtained from the published model) map.

obtained by choosing suitable program parameters. In any case, *FLEX* proved to be a valid alternative to the others.

6. Conclusions

Theoretical results and practical applications described in this paper show that: (*a*) direct methods can solve protein structures without prior information on the heavy-atom structure. However, this information, when available, does not afford supplementary valuable information; (*b*) solvent-flattening procedures are particularly suitable for phase extension and refinement of direct-methods phases. A new procedure (*FLEX*) has proved to be highly competitive with most of the used flattening methods. There are however various limitations of *FLEX*. *E.g.* (*a*) the values for a significant number of variable parameters (V_f, the relative weights M_1, M_2, the initial averaging radius R_i and its decrement *etc.*) have been heuristically chosen. There is no established theoretical development to which to refer, but only our (limited) experience; (*b*) the procedure is rigid: there is no special criterion for judging the best phase improvement. *FLEX*, however, shows that a judicious choice of the parameters (we have used the same 'default' parameters for all the test structures) can further improve modern solvent-flattening procedures.

APPENDIX *A*
The 'cross-over case'

When F_H is large and F_p is small, the so-called 'cross over' may occur. In this case, despite the condition $\Delta > 0$, the sign of F_p is reversed with respect to F_H and $\cos\delta < 0$. Triplets with some reflections showing cross over cannot be correctly estimated by (1). Indeed, in the absence of any information on $|F_H|$, the most probable assumption $\cos\delta > 0$ is always preferred by (1). Once the heavy-atom structure is known, (4) can be applied instead of (1). Suppose that all the three reflections forming the triplet have symmetry-restricted values, to 0 or π, the Δi, $i = 1, 2, 3$, are all positive, the third reflection shows a cross over. Then,

$$\delta_1 = \delta_2 = 0, \quad \delta_3 = \pi,$$

$$A_n = 2[\sigma_3/\sigma_2^{3/2}]_H|(S_1' - R_1')(S_2' - R_2')(S_3' + R_3')|, \quad (13)$$

$$\xi_n = \pi.$$

Unlike (1), the distribution (4) estimates $\Phi = \pi$, in agreement with the traditional isomorphous derivative techniques. It is worthwhile noticing that A_n may be quite large because it involves a sum [*i.e.* $(S_3' + R_3')$] instead of a difference [*i.e.* $(S_3' - R_3')$]. In order to understand how (4) works, let us consider the case in which the third

reflection is a general reflection: then,

$$A_n = 2[\sigma_3/\sigma_2^{3/2}]_H|(S_1' - R_1')(S_2' - R_2') \times [(S_3 \cos\delta_3)^2 + S_3'^2 \sin^2\delta_3]^{1/2}| = 2[\sigma_3/\sigma_2^{3/2}]_H|(S_1' - R_1')(S_2' - R_2') \times [S_3'^2 + R_3'^2 - 2S_3'R_3' \cos\delta_3]^{1/2}|, \quad (14)$$

$$\tan\xi_n = S_3' \cos\delta_3/(S_3' \cos\delta_3 - R_3').$$

We notice that both $(S_3' + R_3')$ in (13) and $[S_3'^2 + R_3'^2 - 2S_3'R_3' \cos\delta_3]^{1/2}$ in (14) coincide with $|E_H|$.

The above consideration suggests that (4) may be more useful than (1) for dealing with the cross-over case but presents a quite dangerous overestimation of the phase reliability. This is an additional confirmation of the criticism we made to (4) in the main text.

The authors gratefully acknowledge many helpful discussions with Dr Doriano Lamba.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

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

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

Cannillo, E., Oberti, R. & Ungaretti, L. (1983). *Acta Cryst.* A**39**, 68–74.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**52**, 43–48.

Cura, V., Krishnaswamy, S. & Podjarny, A. D. (1992). *Acta Cryst.* A**48**, 756–764.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Acta Cryst.* **14**, 1188–1195.

Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.

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

Giacovazzo, C. & Gonzalez-Platas, J. (1995). *Acta Cryst.* A**51**, 398–404.

Giacovazzo, C., Siliqi, D., Cascarano. G., Caliandro, R. & Melidoro, A. (1997). *Acta Cryst.* A**53**, 253–263.

Giacovazzo, C., Siliqi, D. & Gonzalez-Platas, J. (1995). *Acta Cryst.* A**51**, 811–820.

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

Giacovazzo, C., Siliqi, D. & Ralph, A. (1994). *Acta Cryst.* A**50**, 503–510.

Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst.* A**50**, 609–621.

Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *Acta Cryst.* A**51**, 177–188.

Glover, I., Haneef, I., Pitts, J., Woods, S., Moss, D., Tickle, I. & Blundell, T. L. (1983). *Biopolymers*, **22**, 293–304.

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

Hecht, H., Erdmann, H., Park, H., Sprinzl, M. & Schmid, R. D. (1995). *Nature (London) Struct. Biol.* **2**, 1109–1114.

Hecht, H., Sobek, H., Haag, T., Pfeifer, O. & Van Pee, K. H. (1994). *Nature (London) Struct. Biol.* **1**, 532–537.

Klop, E. A., Krabbendam, H. & Kroon, J. (1987). *Acta Cryst.* A**43**, 810–820.

Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 41–46.

Mathews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., De Kok, A. & Hol, W. G. J. (1992). *Science*, **255**, 1544–1550.

Rayment, I. (1983). *Acta Cryst.* A**39**, 102–116.

Read, R. (1986). *Acta Cryst.* A**42**, 140–149.

Rould, M. A., Perona, J. J. & Steitz, T. A. (1992). *Acta Cryst.* A**48**, 751–756.

Sim, G. A. (1960). *Acta Cryst.* **13**, 511–512.

Wang, B. C. (1985). *Methods in Enzymology*, **115**, 90–112.

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