Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 19 June 2002
Accepted 9 September 2002

Crystal structure solution of proteins by direct methods: an automatic procedure for SIR–MIR and SIRAS–MIRAS cases

Carmelo Giacovazzo,[a]* Massimo Ladisa[b] and Dritan Siliqi[a,c]

[a]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, [b]Istituto di Cristallografia, Consiglio Nazionale delle Ricerche, c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and [c]Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana, Albania. Correspondence e-mail: c.giacovazzo@area.ba.cnr.it

In the previous paper [Giacovazzo & Siliqi (2002). *Acta Cryst.* A**58**, 590–597], a probabilistic approach for the SIR–MIR and the SIRAS–MIRAS cases has been described. The mathematical technique is able to take into account the errors arising from measurements, lack of isomorphism and heavy-atom model substructure. An automatic procedure is here described in which the conclusive formulas of that probabilistic approach have been implemented. The procedure has been successfully applied to several test structures: it can automatically provide, starting from the experimental data, high-quality electron-density maps.

© 2002 International Union of Crystallography
Printed in Great Britain – all rights reserved

1. Symbols and abbreviations

For the general notation, the reader should consult the paper by Giacovazzo & Siliqi (2002).

Other notations:

Z_j: atomic number of the jth atom.

$\Sigma_{3d}, \Sigma_{3p}, \Sigma_{3h} = \sum f_j(\mathbf{h}_1) f_j(\mathbf{h}_2) f_j(\mathbf{h}_3)$, where the summation is extended to derivative, native protein and heavy atoms, respectively. As usual for direct-methods applications, we will approximate the ratio $(\Sigma^{3/2}/\Sigma_3)$ by $(\sigma_2^{3/2}/\sigma_3)$, where $\sigma_n = \sum Z_j^n$.

$\Phi_p \equiv \phi_{p1} + \phi_{p2} + \phi_{p3}$: triplet invariant of the native protein. The subscripts pi stand for $p\mathbf{h}_i$, under the condition that $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$. A similar notation holds for the subscripts di.

$\Delta_{\rm iso} = |F_d|_{\rm obs} - |F_p|_{\rm obs}$: isomorphous difference of the native protein.

$D_i(x) = I_i(x)/I_0(x)$, $I_i(x)$ is the modified Bessel function of order i.

The following papers are denoted as papers I–IX, respectively: Giacovazzo *et al.* (1988); Giacovazzo *et al.* (1994); Giacovazzo, Siliqi & Zanzotti (1995); Giacovazzo, Siliqi & González-Platas (1995); Giacovazzo *et al.* (1996); Giacovazzo & Siliqi (1997); Giacovazzo *et al.* (2001); Giacovazzo *et al.* (2002); Giacovazzo & Siliqi (2002).

2. Introduction

The integration between isomorphous replacement techniques and direct methods was initiated by Hauptman (1982). The triplet phase invariant Φ_p of the native protein was estimated *via* the six moduli $R_{p1}, R_{p2}, R_{p3}, R_{d1}, R_{d2}, R_{d3}$. The problem was revisited in paper I. The conclusive probabilistic formula was of von Mises type:

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

where

$$G = 2[\sigma_3/\sigma_2^{3/2}]_p R_{p1} R_{p2} R_{p3} + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_1 \Delta_2 \Delta_3. \qquad (2)$$

$I_0(x)$ is the modified Bessel function of the order zero and

$$\Delta = (|F_d| - |F_p|)/(\Sigma_H)^{1/2} = \Delta_{\rm iso}/(\Sigma_H)^{1/2}$$

is the pseudo-normalized difference (with respect to the heavy-atom structure).

The first term in (2) is often negligible with respect to the second, which may attain large values even for large proteins.

Papers II–VI were devoted to designing, implementing and testing a practical direct procedure (from now on the *O-procedure*) for the solution of protein structures, which may be described in terms of six steps:

Step 1 – Normalization step. Native data are put on an absolute scale, then the derivative data are scaled on the native data by exploiting some properties of the $P(\Delta)$ distribution (see paper III).

Step 2. A certain number of reflections (usually from 800 to 1000; from now on denoted as set NLAR) with large values of R_p and $|\Delta|$ are selected, and triplet invariants calculated according to (1) and (2).

Step 3 – The phasing step. A starting set of phases is generated by a random process: to them a weighted tangent formula is applied and various trials produced.

Step 4 – Identification of the correct solution. As described in paper II, the classical figures of merit MABS, ALFCOMB and

PSICOMB have been modified to exploit the information contained in the experimental Δ values. In particular:

(i)

$$\mathrm{MABS} = \sum_{\mathbf{h}} \alpha_{\mathbf{h}} \Big/ \Big\langle \sum_{\mathbf{h}} \alpha_{\mathbf{h}} \Big\rangle,$$

where

$$\alpha_{\mathbf{h}} = \left\{ \left[\sum_j G_j \sin(\phi_{p\mathbf{k}_j} + \phi_{p\mathbf{h}-\mathbf{k}_j}) \right]^2 + \left[\sum_j G_j \cos(\phi_{p\mathbf{k}_j} + \phi_{p\mathbf{h}-\mathbf{k}_j}) \right]^2 \right\}^{1/2}.$$

The G_j are fixed by (2).

(ii) ALFCOMB, which depends on the ratios

$$(\alpha_{\mathbf{h}} - \langle \alpha_{\mathbf{h}} \rangle)/\sigma_{\alpha\mathbf{h}},$$

where

$$\sigma^2_{\alpha\mathbf{h}} = \tfrac{1}{2} \sum_j G_j [1 + D_2(G_j) - 2D_1^2(G_j)].$$

(iii) PSICOMB: it is calculated by quartering the reflections with small values of R_p and $|\Delta|$, and then by constructing the PSI0 triplets. In particular, PSICOMB depends on the ratios $\alpha'_{\mathbf{h}}/\sigma_{\alpha'_{\mathbf{h}}}$, where

$$\alpha'_{\mathbf{h}} = \left\{ \left[\sum_j G'_j \sin(\phi_{p\mathbf{k}_j} + \phi_{p\mathbf{h}-\mathbf{k}_j}) \right]^2 + \left[\sum_j G'_j \cos(\phi_{p\mathbf{k}_j} + \phi_{p\mathbf{h}-\mathbf{k}_j}) \right]^2 \right\}^{1/2},$$

$$G'_j = 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_{\mathbf{k}_j} \Delta_{\mathbf{h}-\mathbf{k}_j}$$

$$\sigma_{\alpha'_{\mathbf{h}}} = \left(\sum_j G'^2_j \right)^{1/2}.$$

A combined figure of merit (CFOM) integrates the indications arising from the component FOM's. The efficiency of CFOM will be described in §5. It seems worthwhile stating that our FOM's are not based on the Sayre equation but on the supplementary information contained in the experimental Δ's. Their knowledge reduces the mathematical complexity of the problem from the order $[\sigma_3/\sigma_2^{3/2}]_p$ to the order $[\sigma_3/\sigma_2^{3/2}]_H$ (*i.e.* from the order of the size of the protein to the order of the size of a small molecule).

A differential Fourier synthesis is also calculated, to discard trials with high values of the combined figure of merit CFOM but locating atoms on an allowed origin.

Step 5 – Phase extension up to derivative resolution. For each trial solution selected according to step 4, the phase extension up to the derivative resolution is performed. Batches of about 200 reflections chosen in decreasing order of $|\Delta|$ are progressively phased *via* triplets constituted by one reflection belonging to the set NLAR and two reflections belonging to the actual batch.

Step 6. Phase extension up to native resolution *via* solvent flattening techniques (see paper VI).

The *O*-procedure presents advantages and disadvantages with respect to traditional SIR–MIR techniques. The main advantage is the following: while classical techniques require an intermediate step (the location and the refinement of the heavy atoms), the *O*-procedure may automatically provide protein phases directly from the diffraction data, without the prior knowledge of the heavy-atom structure. However, this structure may be easily determined by the *O*-procedure, *via* an *a posteriori* differential Fourier synthesis and a subsequent automatic refinement. The question then arises whether it was possible to improve the efficiency of the triplet invariants by introducing, into the joint probability distribution

$$P(E_{p1}, E_{p2}, E_{p3}, E_{d1}, E_{d2}, E_{d3}),$$

the supplementary information on the heavy-atom structure. The suggestion (Klop *et al.*, 1987; Fortier *et al.*, 1985) of introducing the so-called 'doublet invariants' proved not fruitful (see paper VI). The problem has been solved in paper IX, where the joint probability distribution

$$P(E_{p1}, E_{p2}, E_{p3}, E_{d1}, E_{d2}, E_{d3} | E_{H1}, E_{H2}, E_{H3})$$

was derived. The resulting conditional distribution

$$P(\Phi_p | \{R_{pi}, R_{di}, E_{Hi}, i = 1, 2, 3\}) \qquad (3)$$

is again of von Mises type, but this time its concentration parameter does not contain any term of order $[\sigma_3/\sigma_2^{3/2}]_H$. The application of (3) to practical cases did not lead to phases better than those obtained *via* the *O*-procedure or *via* the classical SIR techniques.

There are three main disadvantages of the *O*-procedure:

(*a*) A prior estimate of $[\sigma_3/\sigma_2^{3/2}]_H$ (and therefore of the number and of the occupancy factors of the heavy atoms) is

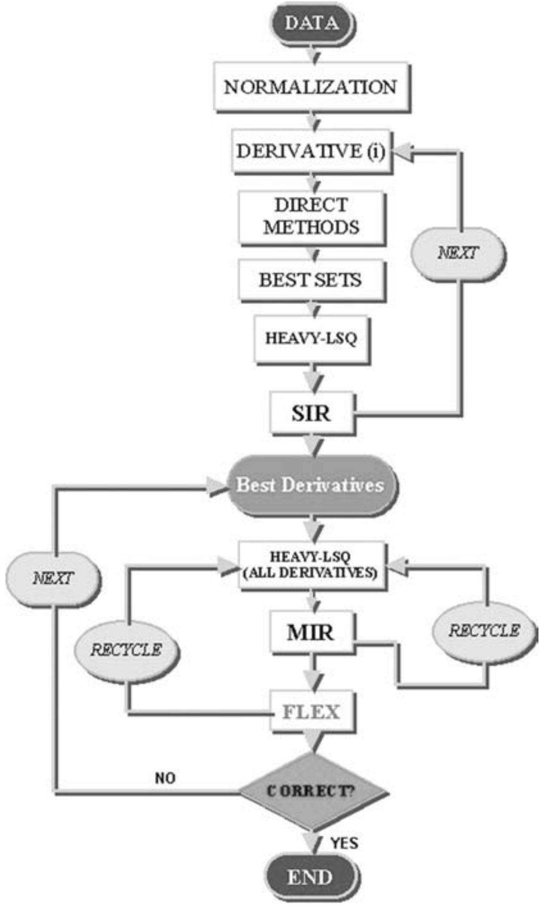

Figure 1
Flow chart of the phasing procedure for the SIR–MIR case.

Table 1
Crystallochemical and diffraction data of the test structures.

NASYM is the number of non-H atoms in the asymmetric unit, RES is the resolution limit of the data (of the native and of each derivative respectively), NREFL is the corresponding value of the measured reflections. The subscripts 1 and 2 to the Pt derivatives of NOX emphasize that the two derivatives were prepared under different conditions.

			Native			Derivative	
Structure code	Space group	NASYM	RES (Å)	NREFL	Heavy atoms	RES (Å)	NREFL
APP[a]	*C*2	302	0.99	17058	Hg	2.00	2108
BPO[b]	*P*$2_1$3	4529	2.35	23956	Au	2.80	15741
					Pt	2.76	7433
DUTPASE[c]	*R*3	1028	1.90	13638	Hg	2.00	11704
					Pt	2.10	9862
E2[d]	*F*432	1853	2.65	10388	Hg	3.00	9179
GLPE[e]	*P*3_2	931	1.06	44798	Ho	2.00	6506
M-FABP[f]	*P*$2_12_12_1$	1101	2.14	7595	Hg	2.18	7125
					Pt	2.15	6586
NOX[g]	*P*4_12_12	1689	2.26	9400	Pt_1	2.26	9068
					Hg	2.59	5425
					Au	2.38	7299
					Pt_2	2.37	6752

References: (*a*) Glover *et al.* (1983); (*b*) Hecht *et al.* (1994); (*c*) Cedergren-Zeppezauer *et al.* (1992); (*d*) Mattevi *et al.* (1992); (*e*) Spallarossa *et al.* (2001); (*f*) Zanotti *et al.* (1992); (*g*) Hecht *et al.* (1995).

necessary to obtain the Δ's from the $\Delta_{\rm iso}$'s. The accuracy of the estimate is not critical and the procedure works well in many cases, but large errors can lower the quality of the protein phases obtained at the end of the process. In these cases, a well refined heavy-atom structure can make classical SIR–MIR techniques more efficient than the *O*-procedure.

(*b*) It is unable to treat errors [*e.g.* equations (1) and (2) were derived by assuming that there is no error in measurements and that no lack of isomorphism occurs].

(*c*) No probabilistic approach was devised to use the triplet invariants for treating the MIR (as well as the SIRAS and the MIRAS) case.

The present paper describes a new procedure (from now on the *N*-procedure) which preserves the advantage of the full automatism and is not affected by any of the disadvantages (*a*)–(*c*). Applications to several practical cases show that the procedure is robust, constitutes a useful tool for macromolecular crystallographers, and is a valuable alternative to different approaches aiming at automated protein phasing (Konnert *et al.*, 1999; Liu *et al.*, 1999; Terwilliger, 1994; Terwilliger & Berendzen, 1996; Vekhter & Miller, 2001; Woolfson *et al.*, 1997).

3. The *N*-procedure

The *N*-procedure for SIR–MIR cases may be described as follows (see the flow chart in Fig. 1):

Step 1 – Normalization of the native and pseudonormalization of all the derivatives (*i.e.* the structure-factor moduli of the derivatives are put on the scale of the native and normalized with respect to it *via* a Wilson plot). Owing to the ignorance of the scattering power of the heavy-atom substructure, the pseudonormalization is a necessary substitute for the normalization.

Step 2 – A statistical test is performed to estimate the scattering power of each heavy-atom substructure. This information is necessary for the use of relation (2). The probabilistic approach and the results of its application to some test structures (see Table 1) are described in §4.

Step 3 – The native and the corresponding derivative structure-factor moduli are settled on the absolute scale (on the basis of the information gained at Step 2).

Step 4 – For the current derivative, Step 3 of the *O*-procedure is performed.

Step 5 – For the current derivative, Step 4 of the *O*-procedure is executed.

Step 6 – For the current derivative, the heavy-atom structure is found and refined by least squares. Then the structure factors F_H are calculated for all the reflections up to the derivative resolution, and the phases φ_p are re-assigned according to equations (20) and (21) of paper IX. This phasing extension process is more efficient than that described at Step 5 of the *O*-procedure (*i.e. via* the tangent formula): the computing time and the penalty to pay in terms of phase error are slightly smaller.

Step 7 – The 'best' derivative is selected: this is the one for which the phasing procedure described at Step 3 of the *O*-procedure is expected to provide the minimum phase error. Steps 8 and following of the *N*-procedure will make clear that a wrong selection would make the crystal structure solution more time consuming. The probabilistic approach used and the results of its applications to our test structures are described in §5.

Step 8 – In the case of MIR, a differential Fourier synthesis $\Delta_{\rm iso}\exp(i\varphi_p)$ is calculated for the other derivatives, and the corresponding heavy-atom substructures are found and refined as at Step 6. The phase values of the protein reflections are then assigned according to equations (37)–(39) of paper IX. The process is repeated until all the available derivatives are considered.

Steps 6–8 are repeated three times. The rationale is the following: the better phase information obtained after the application of the equation (39) is used to improve the estimates of the heavy-atom substructure parameters and consequently the accuracy of the phase estimates.

Step 9 – The solvent flattening procedure *FLEX* (see paper VI) is applied to improve and to extend the phase information up to native resolution.

Step 10 – The quality of the final map is submitted to the user. If the quality of the map is not satisfactory, the next trial with the highest value of PON (see §5) is selected, and the procedure is started again at Step 6.

An optimized procedure for the SIRAS–MIRAS cases should substantially differ from the *N*-procedure described above. For example:

(*a*) The protein triplet invariants should be estimated *via* a formula able to simultaneously exploit isomorphous as well as anomalous differences, instead of *via* (2). This requires the availability of the joint probability distribution function

$$P(E_{p1}, E_{p2}, E_{p3}, E_{d1}^{+}, E_{d1}^{-}, E_{d2}^{+}, E_{d2}^{-}, E_{d3}^{+}, E_{d3}^{-})$$

and the ability of treating errors arising from measurements and from lack of isomorphism. The formula estimating triplet invariants in the SIRAS–MIRAS cases is still not available.

(*b*) The heavy-atom structure refinement should be performed *via* a least-squares process based on the observables $|F_p|$, $|F_d^+|$, $|F_d^-|$, and on the model structure factors F_H^+, F_H^-, rather than on the simplified techniques used at Step 6 of the *N*-procedure.

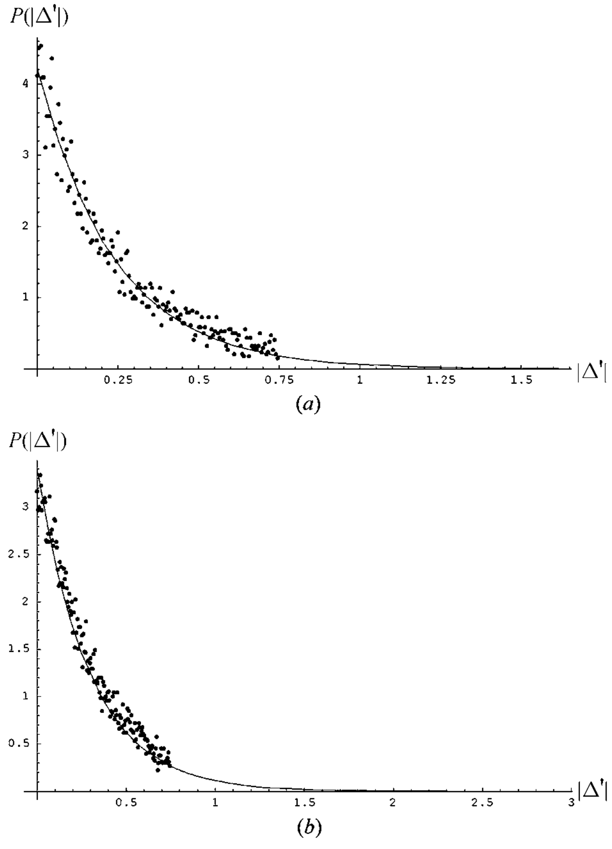

Figure 2
BPO: the experimental $|\Delta'|$ distribution (spots) and the relative best fitting curves (9) for the (*a*) Pt and (*b*) Au derivatives.

In our package, we have implemented a simplified phasing procedure that strictly follows the *N*-procedure but for a few modifications. In particular: (*a*) the values $|F_d^+|$ and $|F_d^-|$ are averaged to $F_d = (|F_d^+| + |F_d^-|)/2$ to simulate the absence of anomalous-dispersion effects. Thus, Steps 1–6 of the *N*-procedure are used without further modifications. (*b*) Step 7 is modified as follows: as soon as the heavy-atom substructure has been refined, the phases ϕ_p are assigned according to equations (46) and (47) of paper IX.

4. Finding the scattering power of the heavy-atom substructure

Hauptman (1982) derived the joint probability distribution function

$$P(R_p, R_d, \phi_p, \phi_d), \tag{4}$$

where $R_p = |E_p|/\Sigma_p^{1/2}$, $R_d = |E_d|/\Sigma_d^{1/2}$. From (4), the marginal distribution

$$P(I, J) = (1-\alpha^2)^{-1/2} \exp\left[-\frac{I+J}{1-\alpha^2}\right] I_0\left[\frac{2\alpha^2}{1-\alpha^2}(IJ)^{1/2}\right] \tag{5}$$

was derived, where I_0 is the modified Bessel function of order zero, and $I = R_p^2$, $J = R_d^2$, $\alpha = (\Sigma_p/\Sigma_d)^{1/2}$.

The change of variable $\Delta = J - I$ transforms $P(I, J)$ into a suitable density $P(I, \Delta)$. The integration over I through the relation

$$\int_0^{+\infty} \mathrm{d}x \exp[-qx] I_0[p(x^2 + 2\gamma x)^{1/2}] = (q^2 + p^2)^{-1/2} \exp\{\gamma[q - (q^2 + p^2)^{1/2}]\} \tag{6}$$

under the condition $q > 0$ leads to the density (Parthasarathy & Srinivasan, 1964)

$$P(|\Delta|) = c^{-1} \exp[-|\Delta|/c], \tag{7}$$

where $c = 1 - \alpha^2 = \Sigma_H/\Sigma_d$.

The distribution (7) is not useful for deriving the scattering power of the heavy-atom structure. Indeed, R_p and R_d are structure-factor moduli normalized with respect to the protein and to the derivative, respectively, and this implies that Σ_d is *a priori* known.

We can however rewrite J as

$$J = |F_d|^2/\Sigma_d = |F_d'|^2(\Sigma_d/\Sigma_p)/\Sigma_d = |F_d'|^2/\Sigma_p, \tag{8}$$

where $|F_d'|^2$ is the value obtained after having rescaled $|F_d|^2$ on the protein scale.

Since the values $|F_d'|^2$ are available from the experiment (*i.e.* it is supposed that the protein data have been put on the absolute scale *via* a Wilson plot), the distribution

$$P(|\Delta'|) = c^{-1} \exp[-|\Delta'|/c] \tag{9}$$

with $\Delta' = J' - I$ may be fitted with the correponding experimental histogram to derive $c = (\Sigma_H/\Sigma_d)^{1/2}$, from which $\Sigma_H/\Sigma_p = (c^{-2} - 1)^{-1}$ may be obtained. Since Σ_p is usually *a*

Table 2
For each test structure and each derivative, the values of $(\Sigma_H + \langle|\mu|^2\rangle)/\Sigma_p$ are listed and compared with the refined values of Σ_H/Σ_p.

Structure code	Heavy atoms	Σ_H/Σ_p	$(\Sigma_H + \langle\lvert\mu\rvert^2\rangle)/\Sigma_p$
APP	Hg	0.077	0.081
BPO	Au	0.028	0.113
	Pt	0.016	0.060
DUTPASE	Hg	0.042	0.067
	Pt	0.038	0.051
E2	Hg	0.021	0.135
GLPE	Ho	0.072	0.136
MFABP	Hg	0.042	0.057
	Pt	0.016	0.035
NOX	Pt_1	0.029	0.032
	Hg	0.085	0.224
	Au	0.067	0.183
	Pt_2	–	0.028

priori known, the estimate of the Σ_H value to introduce into (2) is now available.

The above result is obtained under the hypothesis of the perfect isomorphism. To treat imperfect isomorphism, we can introduce the relation

$$F_d = F_p + F_H + \mu,$$

where μ is a complex number. Under the hypothesis that μ is uncorrelated with F_p and F_H, we obtain

$$\langle|F_d|^2\rangle = \langle|F_p|^2\rangle + \langle|F_H|^2\rangle + \langle|\mu|^2\rangle$$

or, also,

$$\Sigma_d = \Sigma_p + \Sigma_H + \langle|\mu|^2\rangle,$$

from which

$$\frac{\Sigma_d - \Sigma_p}{\Sigma_p} = \frac{\Sigma_H + \langle|\mu|^2\rangle}{\Sigma_p}.$$

It may be argued that suitable statistics over $|\Delta'|$ will provide an estimate of $(\Sigma_H + \langle|\mu|^2\rangle)/\Sigma_p$ rather than of Σ_H/Σ_p.

For two of the test structures quoted in Table 1, we show in Figs. 2 and 3 the experimental distributions of the $|\Delta'|$ values (spots) and the relative best fitting curves (9): such curves closely represent the experimental distributions. A comparison between the estimated $(\Sigma_H + \langle|\mu|^2\rangle)/\Sigma_p$ and the values of Σ_H/Σ_p refined at Step 7 of the *N*-procedure *via* least squares [according to Dickerson *et al.* (1961)] are shown in Table 2. The comparison indicates a realistic agreement between the statistical estimates and the refined values [*i.e.* always $(\Sigma_H + \langle|\mu|^2\rangle)/\Sigma_p > \Sigma_H/\Sigma_p$]. The dash for the second Pt derivative of NOX indicates that no heavy-atom structure model could be obtained *via* least-squares refinement (*i.e.* any model is unstable under refinement).

5. Selecting the best trial solutions and the 'best' derivatives

The automatism of the entire procedure may be secured only if sufficiently 'good' solutions are identified among the various trials produced by direct methods and if a sufficiently good derivative is selected as a pivot of the phase assignment. To check the high efficiency of the figures of merit described in §2, we show in Table 3, for each test structure and for each derivative, the five top ranked (by CFOM) trials and the average phase error $\langle|\Delta\phi|\rangle$ (calculated with respect to the published structure).

Derivatives potentially suitable for our direct phasing process should satisfy the following two conditions: (*a*) the ratio Σ_H/Σ_p should (at any resolution value) be high enough to provide $|\Delta'|$ differences larger than the measurement errors; (*b*) the lack of isomorphism should not provide a dominant contribution to the experimental $|\Delta'|$ values. If both the above conditions are satisfied, a small average phase error (say $\langle|\Delta\phi|\rangle$) is expected when direct procedures are applied. Unfortunately, the statistical results of §4 suggest that the study of the distribution (9) is intrinsically unable to differentiate between 'good' and 'bad' derivatives, because (9) does provide information on the ratio $(\Sigma_H + \langle|\mu|^2\rangle)/\Sigma_p$ but does not discriminate between Σ_H/Σ_p and $\langle|\mu|^2\rangle/\Sigma_p$. Hauptman (1982) suggested the use of the correlation coefficient of the pair (I, J), say

$$\text{CORR} = \frac{\langle(I - \bar{I})(J - \bar{J})\rangle_H}{[\langle(I - \bar{I})^2\rangle_H\langle(J - \bar{J})^2\rangle_H]^{1/2}}, \tag{10}$$

as a tool for discovering the lack of isomorphism. If the isomorphism is imperfect, CORR is expected to be a mono-

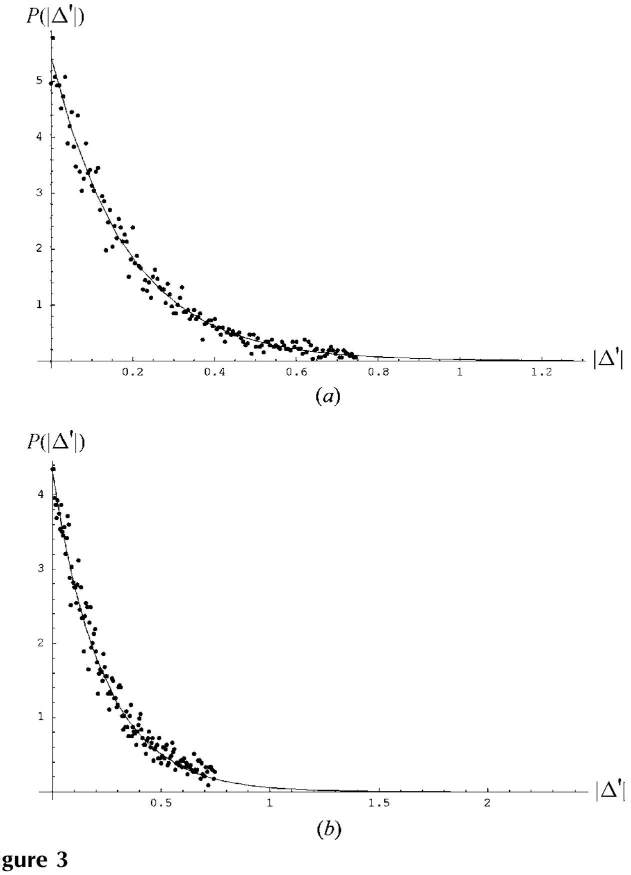

Figure 3
M-FABP: the experimental $|\Delta'|$ distribution (spots) and the relative best fitting curves (9) for the (*a*) Pt and (*b*) Hg derivatives.

Table 3
For each test structure and for each derivative, we show the values of CFOM and the mean phase error for the 'best' (ranked in order of CFOM) five trial solutions (as found among 50 trials).

The good solutions are shown in bold.

Structure code	CFOM ($\langle\lvert\Delta\phi^\circ\rvert\rangle$)				
APP					
Hg	**0.64 (34)**	0.56 (82)	0.35 (84)	0.35 (83)	0.35 (84)
BPO					
Au	**0.79 (32)**	**0.79 (32)**	**0.79 (32)**	0.61 (88)	0.60 (88)
Pt	0.42 (85)	0.42 (85)	0.42 (86)	0.42 (87)	0.42 (88)
DUTPASE					
Hg	**0.59 (34)**	0.50 (49)	0.33 (64)	0.32 (64)	0.28 (66)
Pt	0.42 (60)	**0.21 (52)**	0.16 (71)	0.15 (73)	0.15 (73)
E2					
Hg	**0.79 (29)**	**0.79 (29)**	**0.79 (29)**	0.70 (89)	0.52 (92)
GLPE					
Ho	**0.48 (37)**	0.31 (81)	0.31 (73)	0.31 (80)	0.30 (77)
MFABP					
Hg	**0.46 (40)**	**0.46 (40)**	**0.46 (40)**	0.37 (65)	0.34 (67)
Pt	**0.25 (56)**	**0.24 (54)**	**0.24 (54)**	0.19 (81)	0.18 (86)
NOX					
Pt_1	**0.38 (53)**	**0.38 (53)**	**0.37 (56)**	**0.37 (56)**	0.35 (79)
Hg	**0.49 (56)**	**0.48 (56)**	0.44 (86)	0.44 (86)	0.43 (87)
Au	0.40 (68)	0.39 (68)	0.39 (68)	0.39 (68)	**0.39 (59)**
Pt_2	0.51 (86)	0.49 (88)	0.49 (86)	**0.49 (79)**	0.49 (84)

tonically decreasing function of $\sin\theta/\lambda$. Equation (10) proved unuseful in practice. As an example, in Fig. 4 we depict CORR *versus* $\sin\theta/\lambda$ for the NOX derivatives. The trend is quite similar for Pt_1 and Pt_2: since CORR is nearly constant at the various $\sin\theta/\lambda$ values, both the derivatives should be considered as the best ones. On the contrary, when we apply the direct phasing procedure to the NLAR reflections, the minimum average phase errors are the following:

for Pt_1, $\langle\lvert\Delta\phi\rvert\rangle_{\min} = 53^\circ$, found at trial 1 (the order is defined by CFOM);
for Hg, $\langle\lvert\Delta\phi\rvert\rangle_{\min} = 56^\circ$, found at trial 1 (the order is defined by CFOM);
for Au, $\langle\lvert\Delta\phi\rvert\rangle_{\min} = 59^\circ$, found at trial 5 (the order is defined by CFOM);
for Pt_2, $\langle\lvert\Delta\phi\rvert\rangle_{\min} = 78^\circ$, found at trial 4 (the order is defined by CFOM).

Table 4
Selection of the best derivative for the MIR cases included in the set of our test structures.

Structure code	Heavy atoms	CUL	PO	PON_{best}	ORDTR	$\langle\lvert\Delta\phi^\circ\rvert\rangle_{HR}$
BPO	Au	0.71	1.70	1.22	1	62
	Pt	1.12	1.44	0.61	1	76
DUTPASE	Hg	0.77	1.72	2.23	1	69
	Pt	0.89	1.45	1.37	1	76
MFABP	Hg	0.95	1.46	1.54	1	69
	Pt	0.91	1.33	1.35	1	78
NOX	Pt_1	1.01	1.35	1.34	1	79
	Hg	0.93	1.47	0.94	1	76
	Au	0.93	1.38	1.19	5	89
	Pt_2	1.04	1.26	0.90	4	86

Since both (9) and (10) are unsuitable for selecting the best derivatives, we have used a different approach. Our procedure cannot exploit the quality of the Patterson map, but it is able to directly phase the protein reflections. Therefore it can use figures of merit like the phasing power and/or the Cullis *R* factor to rank the 'goodness' of the derivatives. We use the figure

$$\mathrm{PON} = (\mathrm{PO}/\mathrm{CUL}) * \mathrm{NR},$$

where PO and CUL are the classical phasing power and Cullis *R* factor, respectively, defined by

$$\mathrm{PO} = \sum_n |F_H|^2 \Big/ \sum_n \mathrm{LOC}^2, \quad \mathrm{CUL} = \sum_n \mathrm{LOC} \Big/ \sum_n |\Delta_{\mathrm{iso}}|$$

with $\mathrm{LOC} = ||F_d|_{\mathrm{obs}} - |F_d|_{\mathrm{calc}}|$, n is the number of observed scattering amplitudes and $|F_d|_{\mathrm{calc}} = |F_p + F_H|$. Furthermore, $\mathrm{NR} = n/n_{\max}$, where $n_{\max}$ is the maximum number of observed amplitudes among the various derivatives. Higher values of PON denote better derivatives.

The best derivative is assumed to be the one for which the trial solution with the largest value of PON is obtained (accordingly, the corresponding heavy-atom substructure is considered the most reliable one). In Table 4, we analyse the MIR cases included in the set of our best structures. For each derivative, we show:

(*a*) the best PON value (say $\mathrm{PON}_{\mathrm{best}}$) found among the ten trial solutions with the highest values of CFOM;

(*b*) the order of the trial (ORDTR) providing $\mathrm{PON}_{\mathrm{best}}$;

(*c*) the value $\langle\lvert\Delta\phi\rvert\rangle$ (say $\langle\lvert\Delta\phi\rvert\rangle_{\mathrm{HR}}$) corresponding to such a trial.

Table 4 confirms that the figure of merit CFOM is highly efficient (see the small values of ORDTR); and suggests a high correlation exists between PON and $\langle\lvert\Delta\phi\rvert\rangle_{\mathrm{DM}}$. Accordingly, the default choice of *PROFLEX* is a reasonable one (Au derivative for BPO, Hg derivative for DUTPASE, Hg derivative for M-FABP, Pt_1 derivative for NOX).

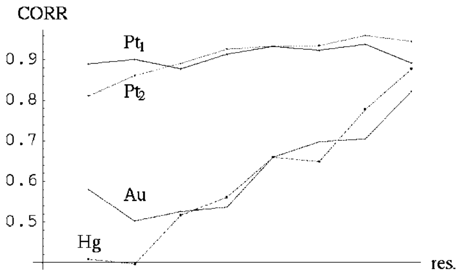

Figure 4
CORR *versus* resolution (Å^{-1}) for NOX.

6. Experimental applications

We have implemented the *N*-procedure in the program *PROFLEX*, a package dedicated to the crystal structure solution of proteins when SIR–MIR or SIRAS–MIRAS cases occur. The goal of this section is to demonstrate that *PROFLEX* is able to automatically derive, from the experimental data and without any user intervention, good quality (*i.e.* perfectly interpretable) electron-density maps. The user is only required to define the atomic species of the heavy atoms:

Table 5
Progress in the phase estimation at various steps of *PROFLEX*.

	Direct methods		SIR/MIR			FLEX			
Structure code	NRDM	$\langle\lvert\Delta\varphi^\circ\rvert\rangle_{\rm DM}$	NRHR	$\langle\lvert\Delta\varphi^\circ\rvert\rangle_{\rm HR}$	CCHR	NRFL	$\langle\lvert\Delta\varphi^\circ\rvert\rangle_{\rm FL}$	CCFL	CPU (min)
APP	497	34	1739	58 (52)	0.46	17058	54 (48)	0.78	9
BPO	837	32	12211	53 (49)	0.46	23956	49 (43)	0.77	49
DUTPASE	960	37	11707	68 (60)	0.48	13638	49 (45)	0.80	27
E2	819	29	6305	55 (50)	0.49	10395	42 (38)	0.87	79
GLPE	960	37	5380	58 (54)	0.42	44798	62 (56)	0.73	74
MFABP	919	40	5228	61 (55)	0.45	7595	46 (38)	0.74	6
NOX	858	53	9303	69 (62)	0.44	9400	54 (45)	0.74	58

the number of symmetry-independent heavy atoms as well as their refined structural parameters are progressively established by *PROFLEX*.

The results of our applications (in default mode) are shown in Table 5. For each test structure, the following information is given:

(*a*) NRDM and $\langle\lvert\Delta\varphi\rvert\rangle_{\rm DM}$ are the number of reflections phased by direct methods and the corresponding average phase error (in degrees), respectively. Both the figures refer to the derivative automatically selected as the 'best'.

(*b*) NRHR and $\langle\lvert\Delta\varphi\rvert\rangle_{\rm HR}$ are the number of reflections phased by the *N*-procedure at the end of the heavy-atom structure refinement (Step 8 of the *N*-procedure) and the corresponding average phase error, respectively (the weighted phase error in parentheses). CCHR is the correlation factor

$$\mathrm{CCHR} = \frac{\langle\rho\rho_{\rm mod}\rangle - \langle\rho\rangle\langle\rho_{\rm mod}\rangle}{[(\langle\rho^2\rangle - \langle\rho\rangle^2)(\langle\rho^2_{\rm mod}\rangle - \langle\rho_{\rm mod}\rangle^2)]^{1/2}}$$

computed between the electron-density map calculated at the end of Step 8 and the electron density corresponding to the published (*i.e.* refined) protein structure (at protein data resolution).

(*c*) NRFL and $\langle\lvert\Delta\varphi\rvert\rangle_{\rm FL}$ are the number of reflections phased by the last run of *FLEX* (Step 9 of the *N*-procedure) and the corresponding phase error, respectively (the weighted phase error in parentheses). CCFL is the correlation factor between the corresponding electron-density map and the map calculated at protein data resolution with published (*i.e.* refined) phases.

(*d*) The CPU time (on a DELL Precision 500, Pentium IV 1.7 GHZ) necessary for *PROFLEX* to provide the final electron-density map.

The results show the progressive gain of phase information along the various stages of the *N*-procedure and demonstrate the efficiency of *PROFLEX* and its ability to perform the entire phasing process in complete automatism. The procedure will soon be made available in a public computer program, combined with other routines able to automatically phase proteins *via* SAD–MAD data.

References

Cedergren-Zeppezauer, E. S., Larsson, G. & Wilson, K. S. (1992). *Nature (London)*, **355**, 740–743.

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.-D. (1988). *Acta Cryst.* A**44**, 45–51.

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

Giacovazzo, C. & Siliqi, D. (2002). *Acta Cryst.* A**58**, 590–597.

Giacovazzo, C. Siliqi, D. & De Caro, L. (2002). *Acta Cryst.* A**58**, 201–207.

Giacovazzo, C., Siliqi, D. & Garcia-Rodriguez, L. (2001). *Acta Cryst.* A**57**, 571–575.

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

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

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

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

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

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

Konnert, J., Karle, J., Karle, I. L., Uma, K. & Baleram, P. (1999). *Acta Cryst.* D**55**, 448–457.

Liu, Y.-D., Gu, Y.-X., Zheng, C.-D., Hao, Q. & Fan, H.-F. (1999). *Acta Cryst.* D**55**, 846–848.

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

Parthasarathy, S. & Srinivasan, R. (1964). *Acta Cryst.* **17**, 1400–1407.

Spallarossa, A., Donahue, J., Larson, T., Bolognesi, M., Bordo, D. (2001). *Structure Fold. Des.* **9**, 1117.

Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 17–23.

Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* D**52**, 749–757.

Vekhter, Y. & Miller, R. (2001). *Acta Cryst.* D**57**, 1048–1051.

Woolfson, M. M., Yao, J.-X. & Fan, H.-F. (1997). *Acta Cryst.* D**53**, 673–681.

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