Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 28 November 2003
Accepted 23 March 2004

© 2004 International Union of Crystallography
Printed in Great Britain – all rights reserved

Locating heavy atoms by integrating direct methods and SIR techniques

Carmelo Giacovazzo,[a,b]* Marat Moustiakimov,[b] Dritan Siliqi[b] and Augusto Pifferi[c]

[a]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, [b]Istituto di Cristallografia, CNR, Via G. Amendola 122/O, 70125 Bari, Italy, and [c]Istituto di Cristallografia, CNR, sez. Monterotondo, Area della Ricerca di Roma, Montelibretti, 00016 Monterotondo St (RM), Italy. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

Direct methods have been applied to the SIR (single isomorphous replacement) case to estimate structure-factor moduli from diffraction magnitudes. The joint probability distribution function $P(E_H, E_p, E_d)$ has been calculated by explicitly considering, as an additional primitive random variable, the cumulative error arising from measurements and from lack of isomorphism. The specific feature of the approach is that it provides estimated values of $|E_H|$ which depend on the experimental diffraction data as well as on errors. Some test structures have been used to check the efficiency of the new estimates. Patterson techniques, using the estimated $|E_H|^2$ values as coefficients, as well as a tangent procedure, have been implemented into a computer program to locate the heavy atoms automatically. All the experimental tests show that the new approach provides results highly competitive with the traditional $|E_H|^2$ estimates.

1. Notation

f_j: scattering factor of the jth atom.
$\sum_p, \sum_d, \sum_H = \sum f_j^2$: where the summation is extended to the protein atoms, to the derivative and to the heavy-atom structure, respectively.
$F_p, F_d, F_H, E_p, E_d, E_H$: F represents the structure factors for protein, derivative and heavy-atom substructure, respectively. E is the corresponding normalized structure factor.
$\Delta_{\rm iso} = (|F_d| - |F_p|)$.

2. Introduction

Traditional soaking methods for obtaining heavy-atom derivatives are a time-consuming and cumbersome process. In recent years, new techniques based on quick-soak methods (Dauter *et al.*, 2000; Sun & Radaev, 2002; Sun *et al.*, 2002), the use of noble gases (Quillin & Matthews, 2002) and innovation in the derivatization strategies (Garman & Murray, 2003) have added new appeal to isomorphous replacement methods.

Once isomorphous diffraction data are available, the following phasing procedure is usually applied: the first step is the determination of the positions of the heavy atoms; the estimate of the native protein phases is obtained in a second step, *via* the combined use of $F_H, |F_p|, |F_d|$ (Blow & Crick, 1959; Terwilliger & Eisenberg, 1987; Terwilliger, 1994).

The contribution of direct methods to the first step started with Steitz (1968), who used only reflections with restricted phase values for solving the carboxy-peptidase substructure, and continued with Neidle (1973), Navia & Sigler (1974), Wilson (1978) and Schevitz *et al.* (1981). The advent of shake-and-bake (Weeks *et al.*, 1994) through the combined use of reciprocal- and real-space refinement made the solution of the substructures more robust. Several well documented programs can be used today for heavy-atom location in single isomorphous replacement (SIR) techniques: *SnB* (Xu *et al.*, 2000), *SHELXD* (Schneider & Sheldrick, 2002), *RANTAN* (Yao, 1981), *ACORN* (Yao, 2002) and *SIR*2002 (Burla *et al.*, 2004).

Hauptman (1982*a*) proposed replacement of the classical two-step procedure by a one-step process: the new approach was able to directly estimate triplet phase invariants from isomorphous diffraction magnitudes *via* the method of joint probability distribution functions. Hauptman's formula was modified by Giacovazzo *et al.* (1988) and applied in a series of papers to a number of practical cases (see references in Giacovazzo *et al.*, 1995). More recently, the method has been optimized (Giacovazzo & Siliqi, 2002; Giacovazzo *et al.*, 2002): the new mathematical technique is able to take the errors into account, *i.e.* the error is a primitive random variable, as well as the atomic coordinates, and has been successfully implemented in a procedure for the automatic crystal structure solution of proteins.

The same mathematical approach (*i.e.* taking errors into account) has been also applied by Giacovazzo *et al.* (2003) to optimize the formulas, previously derived by Hauptman (1982*b*) and by Giacovazzo (1983), which estimate triplet phase invariants in the single-wavelength anomalous diffraction (SAD) case. The applications to real cases were disappointing. The phasing strategy was then changed: instead of using direct methods to estimate triplet phase invariants in the SAD or MAD (multiple-wavelength anomalous diffraction) cases, direct methods were first used to estimate the structure-

factor moduli of the anomalous scatterer substructure, from which the anomalous scatterer positions could be found (Burla *et al.*, 2002, 2003). Then such information was used as prior in a new mathematical approach aiming at finding the protein phases *via* the method of joint probability distribution functions (Giacovazzo & Siliqi, 2004). The new conclusive formulas proved particularly effective for both SAD and MAD cases.

The above results suggest the application of direct methods to estimate, in the SIR case, the structure-factor moduli of the heavy atoms (rather than to estimate triplet phase invariants). This job is the main aim of this paper: the joint probability distribution function $P(E_H, E_p, E_d)$ will be calculated, from which the estimates of $|F_H|$ will be derived, given the diffraction magnitudes. To check the usefulness of the new approach in practical cases, the heavy atoms of some test structures will be located by Patterson techniques and/or by direct methods. It is worthwhile stressing that the estimates of $|F_H|$ obtained in this paper can be easily implemented in all of the most widely used programs for treating isomorphous data (*e.g. SnB*, *SHELXD*, *RANTAN*, *ACORN*, *SIR*2002), and are potentially useful for increasing their efficiency.

3. The SIR case: the estimate of $|F_H|$

We will study the joint probability distribution function

$$P(E_H, E_p, E_d) \tag{1}$$

under the following assumptions:

(*a*) the atomic positions of the native protein structure and the positions of the heavy atoms in the derivative structure are the primitive random variables of our probabilistic approach;

(*b*)

$$|F_d| \exp(i\varphi_d) = |F_p| \exp(i\varphi_p) + |F_H| \exp(i\varphi_H) + |\mu_d| \exp(i\theta_d) \tag{2}$$

is the structure factor of the derivative. It is the sum of the protein structure factor F_p, of the heavy-atom structure factor F_H, and of an eventual error $\mu_d = |\mu_d| \exp(i\theta_d)$ cumulating the effects of the measurement errors and of the lack of isomorphism. The variable μ_d is an additional primitive random variable, for which we assume that $\langle \mu_d \rangle = 0$, and that

$$\langle |\mu_d|^2 \rangle = \langle |\mu_i|^2 \rangle + \langle |\mu_m|^2 \rangle,$$

where $\langle |\mu_m|^2 \rangle$ arises from intensity measurement statistics and $\langle |\mu_i|^2 \rangle$ from the lack of isomorphism. In accordance with (2), and in the absence of correlation among F_p, F_H and μ_d, it is

$$\langle |F_d|^2 \rangle = \textstyle\sum_d = \langle |F_p|^2 \rangle + \langle |F_H|^2 \rangle + \langle |\mu_d|^2 \rangle.$$

The characteristic function of (1) is

$$C(u_H, u_p, u_d, v_H, v_p, v_d) \simeq \exp\big\{-\tfrac{1}{4}[(u_H^2 + v_H^2) + (u_p^2 + v_p^2) + (u_d^2 + v_d^2) + 2k_{13}(u_H u_d + v_H v_d) + 2k_{23}(u_p u_d + v_p v_d)]\big\}.$$

$u_H, u_p, u_d, v_H, v_p, v_d$ are carrying variables associated with

$$A_H = \Big(\sum_{j=1}^{H} f_j \cos 2\pi \mathbf{h}\cdot\mathbf{r}_j\Big) \Big/ \big(\varepsilon \textstyle\sum_H\big)^{1/2},$$

$$A_p = \Big(\sum_{j=1}^{p} f_j \cos 2\pi \mathbf{h}\cdot\mathbf{r}_j\Big) \Big/ \big(\varepsilon \textstyle\sum_p\big)^{1/2},$$

$$A_d = \Big(\sum_{j=1}^{p} f_j \cos 2\pi \mathbf{h}\cdot\mathbf{r}_j + \sum_{j=1}^{H} f_j \cos 2\pi \mathbf{h}\cdot\mathbf{r}_j + |\mu_d| \cos\vartheta_d\Big) \Big/ \big(\varepsilon \textstyle\sum_d\big)^{1/2},$$

$$B_H = \Big(\sum_{j=1}^{H} f_j \sin 2\pi \mathbf{h}\cdot\mathbf{r}_j\Big) \Big/ \big(\varepsilon \textstyle\sum_H\big)^{1/2},$$

$$B_p = \Big(\sum_{j=1}^{p} f_j \sin 2\pi \mathbf{h}\cdot\mathbf{r}_j\Big) \Big/ \big(\varepsilon \textstyle\sum_p\big)^{1/2},$$

$$B_d = \Big(\sum_{j=1}^{p} f_j \sin 2\pi \mathbf{h}\cdot\mathbf{r}_j + \sum_{j=1}^{H} f_j \sin 2\pi \mathbf{h}\cdot\mathbf{r}_j + |\mu_d| \sin\vartheta_d\Big) \Big/ \big(\varepsilon \textstyle\sum_d\big)^{1/2},$$

respectively. $A_H, B_H, A_p, B_p, A_d, B_d$ are the normalized real and imaginary components of E_H, E_p, E_d, respectively,

$$k_{13} = \big(\textstyle\sum_H / \sum_d\big)^{1/2}, \qquad k_{23} = \big(\textstyle\sum_p / \sum_d\big)^{1/2}.$$

We obtain

$$P(A_H, A_p, A_d, B_H, B_p, B_d) = \pi^{-3} (\det \mathbf{K})^{-1/2} \times \exp[-1/2(\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T})], \tag{3}$$

where

$$\overline{\mathbf{T}} = [2^{1/2}A_H, 2^{1/2}A_p, 2^{1/2}A_d, 2^{1/2}B_H, 2^{1/2}B_p, 2^{1/2}B_d],$$

$$\mathbf{K} = \begin{vmatrix} \mathbf{Q} & 0 \\ 0 & \mathbf{Q} \end{vmatrix},$$

$$\mathbf{Q} = \begin{vmatrix} 1 & 0 & \big(\sum_H/\sum_d\big)^{1/2} \\ 0 & 1 & \big(\sum_p/\sum_d\big)^{1/2} \\ \big(\sum_H/\sum_d\big)^{1/2} & \big(\sum_p/\sum_d\big)^{1/2} & 1 \end{vmatrix},$$

$$\det(\mathbf{K}) = [\det(\mathbf{Q})]^2 = \big(\langle |\mu_d|^2 \rangle / \textstyle\sum_d\big),$$

$$\mathbf{K}^{-1} = \begin{vmatrix} \mathbf{Q}^{-1} & 0 \\ 0 & \mathbf{Q}^{-1} \end{vmatrix}.$$

In terms of moduli and phases, (3) may be rewritten as

$$P(R_H, R_p, R_d, \varphi_H, \varphi_p, \varphi_d) \cong \pi^{-3} (\det \mathbf{K})^{-1/2} R_H R_p R_d \exp\{-[\lambda_{11}R_H^2 + \lambda_{22}R_p^2 + \lambda_{33}R_d^2 + 2\lambda_{12}R_H R_p \cos(\varphi_H - \varphi_p) + 2\lambda_{13}R_H R_d \cos(\varphi_H - \varphi_d) + 2\lambda_{23}R_p R_d \cos(\varphi_d - \varphi_p)]\}, \tag{4}$$

where λ_{ij} is the generic element of the matrix $\mathbf{K}^{-1}$. We first apply the approximation $\varphi_p \simeq \varphi_d$ (Giacovazzo & Siliqi, 2002) and then we use standard mathematical techniques to derive, from (4), the following marginal and conditional distributions:

(*a*)

$$P(R_H, R_p, R_d) \cong S R_H R_p R_d \exp\{-[\lambda_{11}R_H^2 + \lambda_{22}R_p^2 + \lambda_{33}R_d^2]\} \times I_o(2R_H X),$$

where $X = \lambda_{12}R_p + \lambda_{13}R_d$ and S is a suitable scale factor;

(*b*)

$$P(R_H|R_p, R_d) \cong LR_H \exp[-\lambda_{11}R_H^2]I_o(2R_HX), \qquad (5)$$

where $L = 2\lambda_{11}\exp(-X^2/\lambda_{11})$.

Equation (5) shows how the *P*1 Wilson distribution $W(R_H) = 2R_H\exp(-R_H^2)$ is modified by the prior knowledge of R_p, R_d,

$$P(R_H|R_p, R_d) \cong W(R_H)M(R_H, \lambda_{11}, X),$$

where

$$M(R_H, \lambda_{11}, X) = \lambda_{11}\exp[(1-\lambda_{11})R_H^2 - X^2/\lambda_{11}]I_o(2R_HX).$$

The function M is the product of two functions: a rapidly decreasing exponential function [$\lambda_{11} > 1$, see (11)] and the monotonic increasing function I_o. Accordingly, the form and the location of P will depend on the λ_{11} and X parameters.

The location may be calculated as follows. Since

$$\int_0^\infty x^\mu \exp(-\alpha x^2)I_o(bx) = \Gamma[(\mu+1)/2][2\alpha^{(\mu+1)/2}]^{-1} \times \exp[b^2/(4\alpha)]\ {}_1F_1\left(\frac{1-\mu}{2}; 1; \frac{-b^2}{4\alpha}\right),$$

where Γ and ${}_1F_1$ are the gamma and the confluent hypergeometric function, respectively, we obtain

$$\langle R_H|R_p, R_d\rangle = 2^{-1}(\pi/\lambda_{11})^{1/2}\ {}_1F_1\left(-\tfrac{1}{2}; 1; -\frac{X^2}{\lambda_{11}}\right).$$

In its turn, ${}_1F_1\left(-\frac{1}{2}; 1; -z^2\right)$ is well approximated by the hyperbole $y = [1 + 2z^2/\pi^{1/2}]^{1/2}$ in the full range $(0, \infty)$: accordingly, the expected value of R_H may be calculated *via* the simpler expression

$$\langle R_H|R_p, R_d\rangle = \tfrac{1}{2}(\pi/\lambda_{11})^{1/2}\left(1 + \frac{4}{\pi}\frac{X^2}{\lambda_{11}}\right)^{1/2}. \qquad (6)$$

The standard deviation σ_{R_H} associated with the estimate (6) may be calculated as follows. Since

$$\langle R_H|R_p, R_d\rangle^2 = \frac{\pi}{4}\lambda_{11}^{-1} + X^2\lambda_{11}^{-2} \qquad (7)$$

and

$$\langle R_H^2|R_p, R_d\rangle = \lambda_{11}^{-1} + X^2\lambda_{11}^{-2}, \qquad (8)$$

then

$$\sigma_{R_H} = [\langle R_H^2|..\rangle - \langle R_H|..\rangle^2]^{1/2} = \left[\left(1 - \frac{\pi}{4}\right)\lambda_{11}^{-1}\right]^{1/2}, \qquad (9)$$

from which

$$\frac{\langle R_H|..\rangle}{\sigma_{R_H}} = \left[\frac{\pi/4 + X^2/\lambda_{11}}{1 - \pi/4}\right]^{1/2}. \qquad (10)$$

The conditional expected values of R_H, R_H^2 and R_H/σ_{R_H} can be calculated by observing that

$$\lambda_{11} = \left(\textstyle\sum_H + \langle|\mu_d|^2\rangle\right)/\langle|\mu_d|^2\rangle, \qquad (11)$$

$$\lambda_{12} = \left(\textstyle\sum_p \sum_H\right)^{1/2}/\langle|\mu_d|^2\rangle, \qquad (12)$$

$$\lambda_{13} = -\left(\textstyle\sum_H \sum_d\right)^{1/2}/\langle|\mu_d|^2\rangle. \qquad (13)$$

In particular, we obtain

$$\langle|F_H|^2\rangle = \frac{\sum_H}{\left(\sum_H + \langle|\mu_d|^2\rangle\right)}\left[\langle|\mu_d|^2\rangle + \frac{\sum_H}{\left(\sum_H + \langle|\mu_d|^2\rangle\right)}\Delta_{\rm iso}^2\right]. \qquad (14)$$

It is worthwhile relating the above result with previous results in the literature. Perutz (1956) approximated $|F_H|^2$ with the difference $\left(|F_d|^2 - |F_p|^2\right)$. Blow & Crick (1959) and Rossmann (1960) suggested a better approximation: $|F_H|^2 \simeq \Delta_{\rm iso}^2$. A deeper analysis was performed by Philips (1966) and Dodson & Vijayan (1971), suggesting the type and the weight of the interatomic vectors in a $\Delta_{\rm iso}^2$ Patterson synthesis. Further weighting schemes have been proposed by Blessing & Smith (1999) and by Grosse-Kunstleve & Brunger (1999).

None of the previous authors derived, *via* a probabilistic approach, the effect of the errors on the evaluation of the moduli $|F_H|^2$. Conversely, (14) suggests that, if $\langle|\mu_d|^2\rangle = 0$, our probabilistic approach confirms the Blow and Rossmann approximation $\langle|F_H|^2\rangle \simeq \Delta_{\rm iso}^2$. If $\langle|\mu_d|^2\rangle \neq 0$, the Blow and Rossmann estimate should be affected by a systematic error, increasing with $\langle|\mu_d|^2\rangle$.

4. The practical estimate of $|E_H|$

The preceding session suggests that the most suitable approximations of R_H can be derived from the right-hand sides of (6) and (8). However, their use requires the simultaneous knowledge of the scattering power of the heavy-atom substructure $\sum_H$ (a quantity necessary to scale $|F_d|$ with respect to $|F_p|$) and an independent estimate of the cumulative error (*i.e.* $\langle|\mu_d|^2\rangle$). Both these quantities are not accessible from the experimental diffraction data: indeed, in the early stages of the phasing process, we ignore both the number of heavy atoms and their occupancies (they define $\sum_H$), and the component μ_i of μ_d arising from the lack of isomorphism.

On the other hand, suitable statistics (Giacovazzo *et al.*, 2002) may be applied to experimental data to estimate the value of $(\sum_H + \langle|\mu_d|^2\rangle)$. We have used this last result to design a procedure for the experimental estimate of the right-hand sides of (6) and (8), which may be schematized as follows:

Step 1: The value of

$$\left(\textstyle\sum_H + \langle|\mu_d|^2\rangle\right) \qquad (15)$$

is estimated. Since the value of $\sum_H$ is ignored, we assume the experimental value of (15) as the best estimate of $\sum_H$.

Step 2: F_p is normalized *via* the Wilson plot procedure, and F_d *via* a differential Wilson plot (Giacovazzo *et al.*, 1994), by using the value of (15) for the differential scaling.

Step 3: The value of $\Delta_{\rm iso}^{n2}$ is calculated as

$$\Delta_{\rm iso}^{n2} = (\Delta_{\rm iso})^2/\left(\textstyle\sum_H + \langle|\mu_d|^2\rangle\right).$$

We now use (11)–(13) to rewrite (7)–(10) in a form depending on the ratio $\sum_H/\langle|\mu_d|^2\rangle$. We obtain

Table 1
Main crystallochemical data of the test structures.

NA is the number of non-H atoms in the asymmetric unit, RES is the resolution limit of the experimental diffraction data (of the native and of each derivative, respectively), NREFL is the corresponding value of measured reflections. Under the heading 'Heavy atoms', the atomic species and the number of heavy atoms in the asymmetric unit are specified.

			Native		Derivative			
Structure code	Space group	NA	RES (Å)	NREFL	RES (Å)	NREFL	Heavy atoms	
BPO†	$P2_13$	4529	2.35	23956	2.80	15741	Au	2
					2.76	7433	Pt	2
DUTPASE‡	$R3$	1028	1.90	13638	2.00	11704	Hg	1
					2.10	9862	Pt	1
E2§	$F432$	1853	2.65	10388	3.00	9179	Hg	1
GLPE¶	$P3_2$	931	1.06	44798	2.00	6506	Ho	2
M-FABP††	$P2_12_12_1$	1101	2.14	7595	2.18	7125	Hg	1
					2.15	6586	Pt	2
NOX‡‡	$P4_12_12$	1689	2.26	9400	2.26	9068	Pt_1	1
					2.59	5425	Hg	3
					2.38	7299	Au	2
					2.37	6752	Pt_2	2

† Hecht *et al.* (1994). ‡ Cedergren-Zeppezauer *et al.* (1992). § Mattevi *et al.* (1992). ¶ Spallarossa *et al.* (2001). †† Zanotti *et al.* (1992). ‡‡ Hecht *et al.* (1995).

$$\langle R_H|R_p, R_d\rangle = \frac{1}{(1+\langle|\sigma_d|^2\rangle^{1/2}}\left[\frac{\pi}{4}\langle|\sigma_d|^2\rangle + \Delta^{n2}_{\rm iso}\right]^{1/2}, \quad (16)$$

$$\langle R^2_H|R_p, R_d\rangle = \frac{1}{(1+\langle|\sigma_d|^2\rangle)}[\langle|\sigma_d|^2\rangle + \Delta^{n2}_{\rm iso}] \quad (17)$$

and

$$\sigma_{R_H} = \left[\left(1-\frac{\pi}{4}\right)\frac{\langle|\sigma_d|^2\rangle}{(1+\langle|\sigma_d|^2\rangle)}\right]^{1/2}, \quad (18)$$

with $\langle|E_H|^2\rangle = \langle|F_H|^2\rangle/\sum_H$, $\Delta^2_{\rm iso}/\Delta^{n2}_{\rm iso} = \Delta^2_{\rm iso}/\left(\sum_H + \langle|\mu_d|^2\rangle\right)$, $\langle|\sigma_d|^2\rangle = \langle|\mu_d|^2\rangle/\sum_H$.

Equation (17) is plotted in Fig. 1 for various values of the parameter $\langle|\sigma_d|^2\rangle$. The corresponding straight line:

(*a*) passes through the origin when $\langle|\sigma_d|^2\rangle = 0$ and bisects the quadrant ($|E_H|^2$, $\Delta^{n2}_{\rm iso}$). In this case, the relation $\langle|E_H|^2\rangle \simeq \Delta^{n2}_{iso}$ provides an unbiased estimate of $|E_H|^2$, whatever the $\Delta^{n2}_{\rm iso}$ value.

(*b*) does not intersect the origin when $\langle|\sigma_d|^2\rangle \neq 0$, and its slope decreases for increasing values of $\langle|\sigma_d|^2\rangle$. As a general trend, the relation $\langle|E_H|^2\rangle \simeq \Delta^{n2}_{iso}$ underestimates $|E_H|^2$ when $\Delta^{n2}_{\rm iso}$ is small and overestimates $|E_H|^2$ when $\Delta^{n2}_{\rm iso}$ is large. The larger the value of $\langle|\sigma_d|^2\rangle$, the larger the overestimation and underestimation effects.

It is clear, however, that the formulas (16)–(18) have to be modified in order to be applied to practical cases. Their new expressions should take into account (*a*) the approximation introduced in Step 1 of the procedure; (*b*) the fact that only the component μ_m of μ_d is experimentally available; (*c*) the fact that $\langle|\sigma_d|^2\rangle$ is not accessible from experiment.

We found that assuming $\langle|\sigma_d|^2\rangle^{1/2} = 10\langle|\mu_m|^2\rangle^{1/2}/(\sum_H + \langle|\mu_d|^2\rangle)^{1/2}$ (the average is calculated per resolution shell) makes (16)–(18) simple and effective tools for estimating $|E_H|$.

We have applied (17) to the test structures quoted in Table 1. In the table, we give the code names, the space group, the number of atoms in the asymmetric unit (water molecules excluded) and, for both the native and the derivative, the data resolution and the number of measured reflections. For each derivative we also quote the species of the heavy atoms and their number in the asymmetric unit.

To estimate the agreement between the $|E_H|$ values estimated *via* (17) and the true values (*i.e.* those calculated using coordinates and occupancy factors in the published structure), we use the residual

$$R_E = \sum_h |\langle|E_H|\rangle - |E_{H\rm true}||/\sum_h |E_{H\rm true}|$$

and we compared its values with

$$R_\Delta = \sum_h ||\Delta^n_{\rm iso}| - |E_{\rm true}||/\sum_h |E_{\rm true}|.$$

The reader may immediately appreciate that $R_E << R_\Delta$ in Table 2 for all the test structures (they were originally solved by SIR–MIR techniques; M-FABP was solved by combining MIR and molecular replacement methods). One can therefore expect that Patterson techniques (see §5) or direct methods

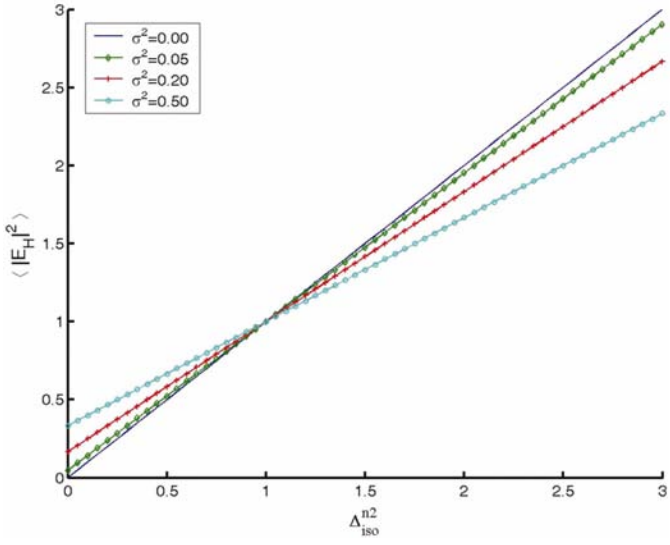

Figure 1
Equation (17) is plotted for various values of the parameter $\langle|\sigma_d|^2\rangle$.

(see §6) will provide better results when (17) is used rather than when the classical Δ_{iso} is employed.

5. Location of the heavy atoms *via* Patterson techniques

A general approach for the automatic determination of the heavy-atom substructure is the so-called *implication function* $I_s(\mathbf{r})$ (Simpson *et al.*, 1965; Pavelčík, 1988; Pavelcik, 1998; Pavelčík *et al.*, 1992; Grosse-Kunstleve & Brunger, 1999). It may be considered as the product of a symmetry operation which transforms a Harker section into a function having maxima at the possible atomic positions compatible with the particular symmetry element generating the Harker section,

$$I_s(\mathbf{r}) = P(\mathbf{r} - \mathbf{C}_s\mathbf{r})/n_s, \tag{19}$$

where $(\mathbf{r} - \mathbf{C}_s\mathbf{r})$ is the Harker vector created by the sth symmetry element $\mathbf{C}_s$, n_s is the multiplicity, equal to the number of symmetry operators, which generates the given Harker vector. For high-symmetry space groups, (19) may be generalized into

$$\mathrm{SMF}(\mathbf{r}) = \min_{s=1}^{\bar{m}} I_s(\mathbf{r}), \tag{20}$$

where the minimum operator Min indicates that the lowest value among the $\bar{m}$ functions I_s has been chosen.

To eliminate spurious maxima, we have calculated the $S(\mathbf{r})$ (*minimum superposition function*) map defined by

$$S(\mathbf{r}) = \mathrm{Min}[P(\mathbf{r} - \mathbf{r}_1), \mathrm{SMF}(\mathbf{r})], \tag{21}$$

where $\mathbf{r}_1$ is a trial atom selected from the peaks in the SMF($\mathbf{r}$) map.

We have applied to the test structures a modified form of the program *SIR*2002 (described by Burla *et al.*, 2004). The procedure is of multisolution type: more positional vectors $\mathbf{r}_1$ may be used and, correspondingly, more trial solutions may be found. For all the test structures, the correct solution has been immediately found by using (15) and the largest peak in the SMF($\mathbf{r}$) map.

In Table 3, we show the results for MFABP obtained *via* the automatic use of (17). We will shortly use its experimental data for checking our theoretical results both for Patterson techniques and for direct-methods phasing. In Table 3, we give, for each derivative:

(*a*) the positions of the highest intensity peaks in the final electron-density maps obtained when Δ_{iso}^{n2} and $\langle|E_H|^2\rangle$ are used; the peaks in bold type correspond to the true heavy-atom positions.

(*b*) the average distance ($\langle$DIST$\rangle$, in Å) between the true atomic positions and the positions found by the program.

It may be observed that, when (17) is used:

(i) the two Pt and the Hg atoms are all more accurately positioned;

(ii) the contrast of the heavy-atom peak intensities with respect to the other peaks is higher.

Similar results are obtained with the other test structures. For example, for GLPE, the two peaks with the highest intensity correspond to the true heavy-atom positions when both Δ_{iso}^{n2} and $\langle|E_H|^2\rangle$ are used. However, the value of $\langle$DIST$\rangle$ is equal to 1.33 Å in the first case, and is equal to 0.19 Å in the second case.

Table 2
For each derivative, the values of R_E (%) and of R_Δ (%) are given.

Structure code	Derivative	R_Δ (%)	R_E (%)
BPO	Au	57	37
	Pt	62	39
DUTPASE	Hg	59	36
	Pt	67	36
E2	Hg	56	39
GLPE	Ho	60	39
M-FABP	Hg	61	36
	Pt	64	42
NOX	Pt_1	69	39
	Hg	69	42
	Au	66	41
	Pt_2	69	43

6. Location of the heavy atoms *via* direct methods

We used a modified form of the program *SIR*2002 to find the heavy-atom substructure *via* tangent procedures. Applications were made according to two protocols:

Protocol 1: $|\Delta_{iso}^{n}|$ is used as an approximation of $|E_H|$;

Protocol 2: the R_H estimate as provided by (17) is employed.

In both protocols, 1000 reflections were submitted to the tangent process, and 60 trials only were explored by a random

Table 3
M-FABP: for each derivative, the coordinates of the largest intensity Fourier peaks are given, as provided by the modified version of the program *SIR*2002.

Columns 2–4 refer to the case in which Δ_{iso}^{n2} is used, columns 5–8 are obtained when $\langle|E|^2\rangle$ is employed. The sites corresponding to true positions are in bold type, OCC is the chemical occupancy of the sites (as in the published paper) and $\langle$DIST$\rangle$ is the average distance in Å between the true atomic positions and the corresponding experimental peaks.

	Δ_{iso}^{n2}			$\langle|E|^2\rangle$			
	Peak positions	Intensity	$\langle$DIST$\rangle$	Peak positions	Intensity	$\langle$DIST$\rangle$	OCC
Pt	**0.7488, 0.3093, 0.9346**	5252	0.173	**0.7498, 0.4772, 0.8989**	7355	0.073	0.6
	0.7492, 0.2492, 0.9500	2260	1.110	**0.2503, 0.3061, 0.9353**	2749	0.378	0.4
	0.2507, 0.4937, 0.8993	2074		0.2497, 0.3064, 0.4326	2185		
	0.7473, 0.4597, 0.2988	1344		0.7305, 0.2548, 0.0474	2115		
Hg	**0.3887, 0.4424, 0.7584**	5957	0.117	**0.8899, 0.4445, 0.2685**	7308	0.084	0.4
	0.2898, 0.3213, 0.7374	1901		0.9068, 0.4008, 0.3034	2164		

Table 4
For each test structure, the type of derivative is specified.

For each protocol, n_f/n is the ratio between the number of heavy atoms found and the number of heavy atoms to find, ⟨DIST⟩ is the average distance (Å) between the heavy-atom positions found by the procedure and the published positions.

		Protocol 1		Protocol 2	
Structure code	Derivative	n_f/n	⟨DIST⟩	n_f/n	⟨DIST⟩
M-FABP	Hg	1/1	0.05	1/1	0.17
	Pt	1/2	0.18	2/2	0.20
NOX	Pt1	1/1	0.48	1/1	0.62
	Hg	–	–	1/3	0.31
	Au	2/2	0.40	2/2	0.38
	Pt2	–	–	1/2	0.60
GLPE	Ho	1/2	0.11	2/2	0.18
E2	Hg	1/1	0.32	1/1	0.22
BPO	Au	2/2	0.04	2/2	0.04
	Pt	2/2	0.18	2/2	0.21
DUTPASE	Hg	1/1	0.06	1/1	0.09
	Pt	1/1	0.07	1/1	0.10

starting procedure. We describe in some detail the results obtained for MFABP, Pt derivative, at the end of the phasing process (*i.e.* when the phase-extension procedures have been applied):

(*a*) for Protocol 1, the best solution (among the 60 explored) shows an average phase error equal to 49° for 4600 phased reflections; the corresponding phase error for the best solution of Protocol 2 is 15° for 4505 reflections.

(*b*) both the heavy atoms are located by Protocol 2, for which ⟨DIST⟩ = 0.20 Å. Only one atom is located by the best solution found *via* Protocol 1, with ⟨DIST⟩ = 0.18 Å.

The results for all the test structures are summarized in Table 4. The reader can verify that the efficiency of Protocol 2 to locate the heavy-atom substructure is markedly greater.

7. Conclusions

We have derived, *via* the method of the joint probability distribution functions, probabilistic estimates of the structure-factor moduli of the heavy-atom substructure, given the moduli of the native and of the derivative. The approach takes into account the errors and provides new formulas that have been successfully applied to some practical cases. Since improved estimates of the moduli $|F_H|$ improve both direct phasing and Patterson map quality, it may be expected that the integration of such estimates into modern packages for protein crystal structure solution from isomorphous data can increase their effectiveness.

References

Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Burla, M. C., Carrozzini, B., Caliandro, R., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Polidori, G. (2004). *J. Appl. Cryst.* **37**, 258–264.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* D**59**, 662–669.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2002). *Acta Cryst.* D**58**, 928–935.
Cedergren-Zeppezauer, E. S., Larsson, G., Nyman, P. D., Dauter, Z. & Wilson, K. S. (1992). *Nature (London)*, **355**, 740–743.
Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.
Dodson & Vijayan (1971). *Acta Cryst.* B**27**, 2402–2411.
Garman, E. & Murray, J. W. (2003). *Acta Cryst.* D**59**, 1903–1913.
Giacovazzo, C. (1983). *Acta Cryst.* A**39**, 585–592.
Giacovazzo, C., Cascarano, G. & Zheng, C.-D. (1988). *Acta Cryst.* A**44**, 45–51.
Giacovazzo, C., Ladisa, M. & Siliqi, D. (2002). *Acta Cryst.* A**58**, 598–604.
Giacovazzo, C., Ladisa, M. & Siliqi, D. (2003). *Acta Cryst.* A**59**, 569–576.
Giacovazzo, C. & Siliqi, D. (2002). *Acta Cryst.* A**58**, 590–597.
Giacovazzo, C. & Siliqi, D. (2004). *Acta Cryst.* D**60**, 73–82.
Giacovazzo, C., Siliqi, D. & Gonzalez Platas, J. (1995). *Acta Cryst.* A**51**, 811–820.
Giacovazzo, C., Siliqi, D. & Spagna, D. (1994). *Acta Cryst.* A**50**, 609–621.
Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). *Acta Cryst.* D**55**, 1568–1577.
Hauptman, H. (1982*a*). *Acta Cryst.* A**38**, 289–294.
Hauptman, H. (1982*b*). *Acta Cryst.* A**38**, 632–641.
Hecht, H., Erdmann, H., Park, H., Sprinzl, M. & Schmid, R. D. (1995). *Nature Struct. Biol.* **2**, 1109–1114.
Hecht, H., Sobek, H., Haag, T., Peifer, O. & Van Pee, K. H. (1994). *Nature Struct. Biol.* **1**, 532–537.
Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., De Kok, A. & Hol, W. G. J. (1992). *Science*, **255**, 1544–1550.
Navia, M. A. & Sigler, P. B. (1974). *Acta Cryst.* A**30**, 706–712.
Neidle, S. (1973). *Acta Cryst.* B**29**, 2645–2647.
Pavelčík, F. (1988). *Acta Cryst.* A**44**, 724–729.
Pavelcik, F. (1998). *J. Appl. Cryst.* **31**, 960–962.
Pavelčík, F., Kuchta, L. & Sivy, J. (1992). *Acta Cryst.* A**48**, 791–796.
Perutz, M. F. (1956). *Acta Cryst.* **9**, 867–873.
Philips, D. C. (1966). In *Advances in Structure Research by Diffraction Methods*, edited by R. Brill & R. Mason. New York: Interscience.
Quillin, M. L. & Matthews, B. W. (2002). *Acta Cryst.* D**58**, 97–103.
Rossmann, M. G. (1960). *Acta Cryst.* **13**, 221–226.
Schevitz, R. W., Podjarny, A. D., Zwick, M., Hughes, J. J. & Sigler, P. B. (1981). *Acta Cryst.* A**37**, 669–677.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Simpson, P. G., Dobrott, R. D. & Lipscomb, W. N. (1965). *Acta Cryst.* **18**, 169–179.
Spallarossa, A., Donahue, J., Larson, T., Bolognesi, M. & Bordo, D. (2001). *Struct. Fold. Des.* **9**, 1117.
Steitz, T. A. (1968). *Acta Cryst.* B**24**, 504–507.
Sun, P. D. & Radaev, S. (2002). *Acta Cryst.* D**58**, 1099–1103.
Sun, P. D., Radaev, S. & Kattah, M. (2002). *Acta Cryst.* D**58**, 1092–1098.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 17–23.
Terwilliger, T. C. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 6–13.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Wilson, K. S. (1978). *Acta Cryst.* B**34**, 1599–1608.
Xu, H., Weeks, C. M., Deacon, A. M., Miller, R. & Hauptman, H. A. (2000). *Acta Cryst.* A**56**, 112–118.
Yao, J. X. (1981). *Acta Cryst.* A**37**, 642–644.
Yao, J. X. (2002). *Acta Cryst.* D**58**, 1941–1947.
Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.