Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 23 April 2002
Accepted 2 December 2002

On integrating the techniques of direct methods and SIRAS: the probabilistic theory of doublets and its applications

Hongliang Xu* and Herbert A. Hauptman

Hauptman-Woodward Medical Research Institute and Department of Structural Biology, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, 73 High Street, Buffalo, NY 14203, USA. Correspondence e-mail: xu@hwi.buffalo.edu

The mathematical formalism of direct methods is here applied to the SIRAS (single-isomorphous replacement combined with anomalous scattering) case. Specifically, the joint probability distribution of three structure factors, which plays the central role in the probabilistic theory of the two-phase structure invariants (doublets), is derived. This distribution leads directly to the conditional probability distribution of the two-phase structure invariants, given the values of selected sets of magnitudes. Furthermore, a probabilistic formula for estimating individual phases of the derivative structure is derived, provided that the heavy-atom substructure is assumed to be known. The formulas were tested for experimental SIRAS data and results are reported.

1. Introduction

The integration of direct methods with macromolecular crystallographic techniques *via* probabilistic theory was initiated by Hauptman in 1982. He developed the probabilistic theory of two-phase (doublet) and three-phase (triplet) structure invariants in the SIR (single-derivative isomorphous replacement) and SAS (single-wavelength anomalous scattering) cases (Hauptman, 1982*a*,*b*). In his approach, the neighborhood principle was applied to a pair of isomorphous structures and the joint probability distributions of one doublet and eight triplets were derived. The same distributions were later obtained from the theory of representations (Giacovazzo, 1983; Giacovazzo *et al.*, 1988). In the past 20 years, great progress has been made in the probabilistic approach to (i) improve the estimates of the cosine values of the triplets by incorporating heavy-atom information into Hauptman's formulas (Fortier *et al.* 1985, 1986; Furey *et al.* 1990); (ii) resolve the phase ambiguity arising from the SIR or SAS case by incorporating phase-doublet information into the triplet probability formula (Fan *et al.*, 1984); (iii) extend probabilistic theory to the MAD (multiwavelength anomalous scattering) case (Giacovazzo & Siliqi, 2001).

SIRAS (single isomorphous replacement combined with anomalous scattering) plays a very important role in macromolecular structure determination owing to simultaneous knowledge of $\Delta_{\rm ano}$ and $\Delta_{\rm iso}$, which in theory allows one to solve the phase problem *via* an algebraic approach (Kartha, 1975). However, the probabilistic approach provides robust tools to handle errors arising from experimental SIRAS data measurements, imperfect isomorphism and heavy-atom substructure determination, *etc*. In this paper, we develop the probabilistic theory in the SIRAS case. In particular, we develop the probabilistic formula to estimate individual phase values by exploiting the probabilistic relationship among three doublets arising in the SIRAS case. Such a relationship does not exist in the SIR or SAS case because only one doublet exists in those cases. Let us assume that $\mathcal{F}$ and $\mathcal{G}$ are a pair of isomorphous structures. Structures $\mathcal{F}$ and $\mathcal{G}$ are said to be isomorphous if the atomic position vectors of $\mathcal{F}$ coincide with those of $\mathcal{G}$, and some of the atoms of $\mathcal{F}$ (or $\mathcal{G}$) may have atomic scattering factors equal to zero [corresponding to vacancies in $\mathcal{F}$ (or $\mathcal{G}$)]. In the SIRAS case, normal diffraction data are available for $\mathcal{F}$ and anomalous scattering diffraction data are available for $\mathcal{G}$. One example of the SIRAS case is that $\mathcal{F}$ is a sulfur-containing native protein and $\mathcal{G}$ is obtained from $\mathcal{F}$ by replacing some (or all) of the sulfur atoms in $\mathcal{F}$ by selenium atoms. One special case is that $\mathcal{F}$ is a known heavy-atom substructure of $\mathcal{G}$ (whence $\mathcal{F}$ and $\mathcal{G}$ are isomorphous). The realization of this goal is particularly important since a powerful method for determining the heavy-atom substructure of a macromolecule by combining *Shake-and-Bake* phasing (Weeks *et al.*, 1994) with single-wavelength anomalous diffraction data, even at a resolution of 3–4 Å, has recently been elucidated (*e.g.* Deacon *et al.*, 2000). At least one substructure having as many as 160 selenium atoms has been routinely determined in this way (Delft, personal communication), and it now appears that much larger substructures will also be readily solvable. This work strongly suggests that the ability to develop the technique of direct methods in the SIRAS case would provide an improved method for phase estimation.

In this paper, the joint probability distribution of three structure factors, which plays the central role in the derivation

of the probabilistic estimates of the three doublets, is obtained. Owing to their extreme length, details of the derivation are deposited; only major formulas are given explicitly here. This joint probability distribution leads directly to the major results of this paper: (i) the conditional probability distribution of the two-phase structure invariants (doublets), given selected magnitudes; (ii) the conditional probability distribution of a single phase when a heavy-atom substructure is known. The formulas were tested using experimental SIRAS data and results are reported in §5.

2. The probabilistic theory of the doublets

For an arbitrary pair of isomorphous structures $(\mathcal{F}, \mathcal{G})$, we assume that normal diffraction intensities are available for $\mathcal{F}$, and single-wavelength anomalous scattering data are available for $\mathcal{G}$. In the absence of anomalous scatterers, the normalized structure factor E_H of $\mathcal{F}$ is defined by

$$E_H = |E_H| \exp(i\varphi_H) = (1/\alpha_H^{1/2}) \sum_{j=1}^{N} f_{jH} \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \quad (1)$$

where f_{jH} is the (real) normal atomic scattering factor of the atom labeled j, $\mathbf{r}_j$ is the position vector of the jth atom, N is the number of atoms in the unit cell and

$$\alpha_H = \sum_{j=1}^{N} f_{jH}^2. \quad (2)$$

In the presence of anomalous scatterers, the normalized structure factor G_H of $\mathcal{G}$ is defined by

$$G_H = |G_H| \exp(i\psi_H) = (1/\beta_H^{1/2}) \sum_{j=1}^{N} |g_{jH}| \exp[i(\delta_{jH} + 2\pi \mathbf{H} \cdot \mathbf{r}_j)], \quad (3)$$

where $g_{jH} = g_{jH}^0 + g_j' + ig_j''$ is the (in general complex) atomic scattering factor of the atom labeled j, g_{jH}^0 is the (real) atomic scattering factor in the absence of anomalous scattering, g_j' and g_j'' are the real and imaginary dispersion corrections, $|g_{jH}| = [(g_{jH}^0 + g_j')^2 + (g_j'')^2]^{1/2}$, $\delta_{jH} = \tan^{-1}[g_j''/(g_{jH}^0 + g_j')]$, and

$$\beta_H = \sum_{j=1}^{N} |g_{jH}|^2. \quad (4)$$

Replacing $\mathbf{H}$ by $\bar{\mathbf{H}}$ in (3), we have

$$G_{\bar{H}} = |G_{\bar{H}}| \exp(i\psi_{\bar{H}}) = (1/\beta_H^{1/2}) \sum_{j=1}^{N} |g_{jH}| \exp[i(\delta_{jH} - 2\pi \mathbf{H} \cdot \mathbf{r}_j)]. \quad (5)$$

Owing to the presence of the anomalous scatterers, Friedel's law does not hold and $|G_H|$ and $|G_{\bar{H}}|$, ψ_H and $-\psi_{\bar{H}}$ are in general distinct. In what follows, the atomic scattering factors f_{jH} and g_{jH}, hence α_H and β_H, are presumed to be known.

It will be assumed throughout this paper that the atomic position vectors $\mathbf{r}_j$ are primitive random variables that are uniformly and independently distributed in the unit cell. Thus, the three structure factors E_H, G_H and $G_{\bar{H}}$, as functions of the primitive random variables $\mathbf{r}_j$, are themselves random variables. Owing to the breakdown of Friedel's law, there are now three distinct two-phase structure invariants (the doublets):

$$\omega_1 = \varphi_H - \psi_H, \quad (6)$$

$$\omega_2 = \varphi_H + \psi_{\bar{H}}, \quad (7)$$

$$\omega_3 = \psi_H + \psi_{\bar{H}}. \quad (8)$$

It is worth pointing out that only ω_1 exists in the SIR case and only ω_3 exists in the SAS case. Therefore, the doublet theory developed in this paper has no analog in the SIR or SAS case.

The first neighborhood of each of the doublets (6)–(8) is defined to consist of the three magnitudes

$$R = |E_H|, \quad S = |G_H|, \quad \bar{S} = |G_{\bar{H}}| \quad (9)$$

and the three phases

$$\varphi = \varphi_H, \quad \psi = \psi_H, \quad \bar{\psi} = \psi_{\bar{H}}. \quad (10)$$

For non-negative integers m and n, define

$$C_{mn} = \frac{1}{(\alpha_H^m \beta_H^n)^{1/2}} \sum_{j=1}^{N} f_{jH}^m |g_{jH}|^n \cos(n\delta_{jH}), \quad (11)$$

$$S_{mn} = \frac{1}{(\alpha_H^m \beta_H^n)^{1/2}} \sum_{j=1}^{N} f_{jH}^m |g_{jH}|^n \sin(n\delta_{jH}). \quad (12)$$

Then, X_{mn} and ξ_{mn} are uniquely defined by:

$$X_{mn} \cos \xi_{mn} = C_{mn}, \quad X_{mn} \sin \xi_{mn} = -S_{mn}. \quad (13)$$

Hence, in all that follows, C_{mn}, S_{mn}, X_{mn} and ξ_{mn} are presumed to be known.

Denote by $P(R, S, \bar{S}; \varphi, \psi, \bar{\psi})$ the joint probability distribution of the magnitudes R, S, $\bar{S}$ and the phases φ, ψ, $\bar{\psi}$ of the normalized structure factors E_H, G_H, $G_{\bar{H}}$. Then the joint probability distribution is given by the sixfold integral (Karle & Hauptman, 1958)

$$\begin{aligned} P(R, S, \bar{S}; \varphi, \psi, \bar{\psi}) &= [1/(2\pi)^6] RS\bar{S} \int_{\sigma,\rho,\bar{\rho}=0}^{\infty} \int_{\gamma,\theta,\bar{\theta}=0}^{2\pi} \sigma\rho\bar{\rho} \exp\{-i[R\sigma\cos(\gamma - \varphi) \\ &\quad + S\rho\cos(\theta - \psi) + \bar{S}\bar{\rho}\cos(\bar{\theta} - \bar{\psi})]\} \\ &\quad \times \prod_{j=1}^{N} q_j(\sigma, \rho, \bar{\rho}, \gamma, \theta, \bar{\theta})\, \mathrm{d}\sigma\, \mathrm{d}\rho\, \mathrm{d}\bar{\rho}\, \mathrm{d}\gamma\, \mathrm{d}\theta\, \mathrm{d}\bar{\theta}, \end{aligned} \quad (14)$$

where

$$\begin{aligned} q_j(\sigma, \rho, \bar{\rho}, \gamma, \theta, \bar{\theta}) = \langle \exp\{ & i(f_{jH}\alpha_H^{-1/2})\sigma\cos(2\pi\mathbf{H}\cdot\mathbf{r}_j - \gamma) \\ & + i(|g_{jH}|\beta_H^{-1/2})[\rho\cos(\delta_{jH} + 2\pi\mathbf{H}\cdot\mathbf{r}_j - \theta) \\ & + \bar{\rho}\cos(\delta_{jH} - 2\pi\mathbf{H}\cdot\mathbf{r}_j - \bar{\theta})]\}\rangle_{\mathbf{r}_j}. \end{aligned}$$

Appendix A[1] contains some preliminary formulas, Appendix B the derivation of q_j and $\prod_{j=1}^{N} q_j$, and Appendix C the evaluation of the sixfold integral (14). The final formula, taken from Appendix C, is simply

[1] Appendices A–C are available from the IUCr electronic archive (Reference: DR0028). Services for accessing these data are described at the back of the journal.

$$P(R, S, \bar{S}; \varphi, \psi, \bar{\psi}) = \frac{RS\bar{S}}{\pi^3(1-X_{11}^2)(1-Z^2)} \times \exp\left[-\frac{(1-X_{02}^2)R^2 + (1-X_{11}^2)(S^2+\bar{S}^2)}{(1-X_{11}^2)(1-Z^2)}\right] \times \exp[W_1 RS\cos(\varphi-\psi+\lambda) + W_1 R\bar{S}\cos(\varphi+\bar{\psi}-\lambda) + W_2 S\bar{S}\cos(\psi+\bar{\psi}+\mu)], \quad (15)$$

where X_{11}, X_{02} are defined by (13),

$$Z^2 = [C_{11}^2 + S_{11}^2 + C_{02}^2 + S_{02}^2 - 2C_{11}^2 C_{02} + 2S_{11}^2 C_{02} - 4C_{11}S_{11}S_{02}](1-C_{11}^2-S_{11}^2)^{-1}, \quad (16)$$

$$U\cos\lambda = C_{11} - C_{11}C_{02} - S_{11}S_{02}, \quad (17)$$

$$U\sin\lambda = S_{11} + S_{11}C_{02} - C_{11}S_{02}, \quad (18)$$

$$V\cos\mu = C_{02} - C_{11}^2 + S_{11}^2, \quad (19)$$

$$V\sin\mu = 2C_{11}S_{11} - S_{02}, \quad (20)$$

and

$$W_1 = \frac{2U}{(1-X_{11}^2)(1-Z^2)}, \quad (21)$$

$$W_2 = \frac{2V}{(1-X_{11}^2)(1-Z^2)}. \quad (22)$$

Hence, the joint probability distribution $P(R, S, \bar{S}; \varphi, \psi, \bar{\psi})$ [equation (15)] depends on parameters X_{11}, X_{02}, Z, W_1, W_2, λ and μ, all of which are in turn defined by (16)–(22) in terms of C_{mn} and S_{mn}, each of which is expressed by means of (11)–(12) in terms of the known atomic scattering factors. Hence, for each fixed **H**, the distribution (15) depends on parameters, all of which may be presumed to be known.

If we allow $C_{11} = S_{11} = 0$, then the joint distribution (15) reduces to

$$P(S, \bar{S}; \psi, \bar{\psi}) = \frac{S\bar{S}}{\pi^2(1-X_{02}^2)}\exp\left[-\frac{S^2+\bar{S}^2}{1-X_{02}^2} + \frac{2S\bar{S}X_{02}}{1-X_{02}^2} \times \cos(\psi+\bar{\psi}+\xi_{02})\right]. \quad (23)$$

Equation (23) coincides with the joint distribution of the two-phase structure invariants in the SAS case obtained by Hauptman (1982*b*).

3. The conditional probability distribution without heavy-atom substructure information

With lack of heavy-atom substructure information, we assume that $\mathcal{F}$ is the native protein and $\mathcal{G}$ is the derivative structure. Then $\mathcal{F}$ and $\mathcal{G}$ are a pair of isomorphous structures, and the joint probability distribution [equation (15)] can be applied directly. Specifically, three magnitudes $R, S, \bar{S}$ are known, three phases φ, ψ and $\bar{\psi}$ are unknown.

Table 1
Basic information of test data sets.

	AXE II	AdoHcy
Protein atoms	1442	6787
Heavy atoms	4 Iodine	30 Selenium
Space group	$P2_12_12_1$	$C222$
Resolution (Å)	2.0	2.8
f', f''	−6.70, 4.50	−7.35, 5.92
a, b, c	34.95, 61.05, 72.61	91.93, 168.02, 133.77
Native data	8501	26211
Derivative data	15957	50466

3.1. The conditional probability distribution of $\omega_1 = \varphi - \psi$

Denote by $P(\omega_1|R, S, \bar{S})$ the conditional probability distribution of the random variable $\omega_1 = \varphi - \psi$, given $R, S, \bar{S}$. Then $P(\omega_1|R, S, \bar{S})$ is derived from (15) by fixing $R, S, \bar{S}$, integrating with respect to $\bar{\psi}$ from 0 to 2π, and multiplying by a suitable normalizing constant. The final formula, the second major result of this paper, is

$$P(\omega_1|R, S, \bar{S}) = [2\pi I_0(A_1)]^{-1}\exp[A_1\cos(\omega_1 - \eta_1)], \quad (24)$$

where A_1 and η_1 are defined by

$$A_1\cos\eta_1 = W_1 RS\cos\lambda + 2T(W_1 R\bar{S})T(W_2 S\bar{S})\cos(\lambda+\mu), \quad (25)$$

$$A_1\sin\eta_1 = -W_1 RS\sin\lambda + 2T(W_1 R\bar{S})T(W_2 S\bar{S})\sin(\lambda+\mu), \quad (26)$$

with $T(z) = I_1(z)/I_0(z)$ being the ratio of two modified Bessel functions. Equation (24) implies that the expected value of $\cos(\omega_1 - \eta_1)$ is $T(A_1)$ and, in the favorable case that A_1 is large,

$$\omega_1 = \varphi - \psi \approx \eta_1. \quad (27)$$

3.2. The conditional probability distribution of $\omega_2 = \varphi + \bar{\psi}$

Denote by $P(\omega_2|R, S, \bar{S})$ the conditional probability distribution of the random variable $\omega_2 = \varphi + \bar{\psi}$, given $R, S, \bar{S}$. Then, $P(\omega_2|R, S, \bar{S})$ is derived from (15) by fixing $R, S, \bar{S}$, integrating with respect to ψ from 0 to 2π and multiplying by a suitable normalizing constant. The final formula, the third major result of this paper, is

$$P(\omega_2|R, S, \bar{S}) = [2\pi I_0(A_2)]^{-1}\exp[A_2\cos(\omega_2 - \eta_2)], \quad (28)$$

where A_2 and η_2 are defined by

$$A_2\cos\eta_2 = W_1 R\bar{S}\cos\lambda + 2T(W_1 RS)T(W_2 S\bar{S})\cos(\lambda+\mu), \quad (29)$$

$$A_2\sin\eta_2 = W_1 R\bar{S}\sin\lambda - 2T(W_1 RS)T(W_2 S\bar{S})\sin(\lambda+\mu). \quad (30)$$

Equation (28) implies that the expected value of $\cos(\omega_2 - \eta_2)$ is $T(A_2)$ and, in the favorable case that A_2 is large,

$$\omega_2 = \varphi + \bar{\psi} \approx \eta_2. \quad (31)$$

Table 2
Average value of concentration parameter A_m, $\langle A_m \rangle$, and its corresponding average magnitude of the doublet error, $\langle |\omega_m - \eta_m| \rangle$, using experimental SIRAS data from proteins (*a*) AXE II and (*b*) AdoHcy.

All reflections are sorted in descending order of values of the A_m's and partitioned into eight equal groups.

Group No.	Reflections in group	$\langle A_1 \rangle$	$\langle \lvert\omega_1 - \eta_1\rvert \rangle$ (°)	$\langle A_2 \rangle$	$\langle \lvert\omega_2 - \eta_2\rvert \rangle$ (°)	$\langle A_3 \rangle$	$\langle \lvert\omega_3 - \eta_3\rvert \rangle$ (°)
(*a*) AXE II							
1	1000	3.87	3.7	3.89	3.7	4.99	3.7
2	1000	2.23	5.3	2.23	5.3	2.75	1.5
3	1000	1.55	6.4	1.55	6.4	1.90	1.6
4	1000	1.09	7.1	1.09	7.2	1.36	1.8
5	1000	0.77	9.4	0.77	9.3	0.98	2.5
6	1000	0.53	10.0	0.53	10.5	0.68	2.9
7	1000	0.33	13.9	0.34	13.3	0.44	3.7
8	1218	0.17	20.7	0.17	21.0	0.23	4.6
(*b*) AdoHcy							
1	3340	5.87	2.1	5.87	2.2	5.84	2.3
2	3340	2.65	2.9	2.65	3.0	2.65	3.1
3	3340	1.69	3.6	1.68	3.7	2.04	3.8
4	3340	1.09	4.4	1.09	4.5	1.48	4.5
5	3340	0.67	5.3	0.67	5.4	1.09	5.5
6	3340	0.39	6.9	0.39	7.0	0.80	7.0
7	3340	0.19	10.0	0.19	9.9	0.57	9.9
8	3343	0.05	26.2	0.05	26.5	0.39	21.6

3.3. The conditional probability distribution of $\omega_3 = \psi + \bar{\psi}$

Denote by $P(\omega_3|R, S, \bar{S})$ the conditional probability distribution of the random variable $\omega_3 = \psi + \bar{\psi}$, given R, S, $\bar{S}$. Then, $P(\omega_3|R, S, \bar{S})$ is derived from (15) by fixing R, S, $\bar{S}$, integrating with respect to φ from 0 to 2π and multiplying by a suitable normalizing constant. The final formula, the fourth major result of this paper, is

$$P(\omega_3|R, S, \bar{S}) = [2\pi I_0(A_3)]^{-1} \exp[A_3 \cos(\omega_3 - \eta_3)], \quad (32)$$

where A_3 and η_3 are defined by

$$A_3 \cos\eta_3 = W_2 S\bar{S} \cos\mu + 2T(W_1 RS)T(W_1 R\bar{S}) \cos(2\lambda), \quad (33)$$

$$A_3 \sin\eta_3 = -W_2 S\bar{S} \sin\mu + 2T(W_1 RS)T(W_1 R\bar{S}) \sin(2\lambda). \quad (34)$$

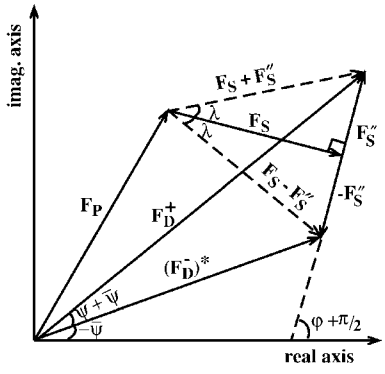

Figure 1
Geometry relationship among F_P, F_D^+, $(F_D^-)^*$, F_S and F''_S, where F_P is the structure factor of the native protein, F_D^+ and $(F_D^-)^*$ are the structure factors of the derivative structure, F_S and F''_S are the contributions to the structure factor of the heavy-atom substructure arising from the real and imaginary components of the anomalous scattering, respectively.

Equation (32) implies that the expected value of $\cos(\omega_3 - \eta_3)$ is $T(A_3)$ and, in the favorable case that A_3 is large,

$$\omega_3 = \psi + \bar{\psi} \approx \eta_3. \quad (35)$$

Again, it should be stressed that the conditional probability distributions of the three doublets, ω_1, ω_2 and ω_3, given the values R, S and $\bar{S}$ of the three magnitudes $|E_H|$, $|G_H|$ and $|G_{\bar{H}}|$, depend not only on R, S and $\bar{S}$ but on the atomic scattering factors, presumed to be known.

4. The conditional probability distribution with heavy-atom substructure information

When heavy-atom substructure information is available, we assume that $\mathcal{G}$ is the derivative structure and $\mathcal{F}$ is the known heavy-atom substructure of $\mathcal{G}$; hence, $\mathcal{F}$ and $\mathcal{G}$ are a pair of isomorphous structures. Specifically, three magnitudes R, S, $\bar{S}$ and one phase φ are known; only ψ and $\bar{\psi}$ are unknowns. Denote by $P(\psi|R, S, \bar{S}; \varphi)$ and $\bar{P}(\bar{\psi}|R, S, \bar{S}; \varphi)$ the conditional probability distributions of the random variable ψ or $\bar{\psi}$, respectively, given R, S, $\bar{S}$ and φ. Then, $P(\psi|R, S, \bar{S}; \varphi)$ or $\bar{P}(\bar{\psi}|R, S, \bar{S}; \varphi)$ is derived from (15) by fixing R, S, $\bar{S}$ and φ, integrating with respect to $\bar{\psi}$ or ψ from 0 to 2π, respectively, and multiplying by a suitable normalizing constant. The final formula, the fifth and sixth major results of this paper, are

$$P(\psi|R, S, \bar{S}; \varphi) = [2\pi I_0(Q)]^{-1} \exp[Q \cos(\psi - \alpha)] \quad (36)$$

and

$$\bar{P}(\bar{\psi}|R, S, \bar{S}; \varphi) = [2\pi I_0(\bar{Q})]^{-1} \exp[\bar{Q} \cos(\bar{\psi} - \bar{\alpha})] \quad (37)$$

with Q, α, $\bar{Q}$ and $\bar{\alpha}$ defined by

$$Q \cos\alpha = W_1 RS \cos(\varphi + \lambda) + 2T_2 T_3 \cos(\varphi - \lambda - \mu), \quad (38)$$

$$Q \sin\alpha = W_1 RS \sin(\varphi + \lambda) + 2T_2 T_3 \sin(\varphi - \lambda - \mu), \quad (39)$$

Table 3
Average values of Q, $\langle Q\rangle$, and the average phase errors, $\langle|\psi-\alpha|\rangle$, using experimental SIRAS data from proteins AXE II and AdoHcy.

All reflections are sorted in descending order of values of the Q's and cumulated in eight groups.

	AXE II			AdoHcy		
Group No.	Reflections in group	$\langle Q\rangle$	$\langle\|\psi-\alpha\|\rangle$ (°)	Reflections in group	$\langle Q\rangle$	$\langle\|\psi-\alpha\|\rangle$ (°)
1	1000	2.06	30.1	3000	1.90	47.1
2	2000	1.84	43.3	6000	1.78	48.6
3	3000	1.71	48.1	9000	1.71	53.4
4	4000	1.61	50.8	12000	1.65	55.1
5	5000	1.52	55.1	15000	1.58	58.7
6	6000	1.44	56.6	18000	1.50	60.2
7	7000	1.34	58.4	21000	1.40	62.2
8	8000	1.22	61.3	24000	1.28	64.3

and

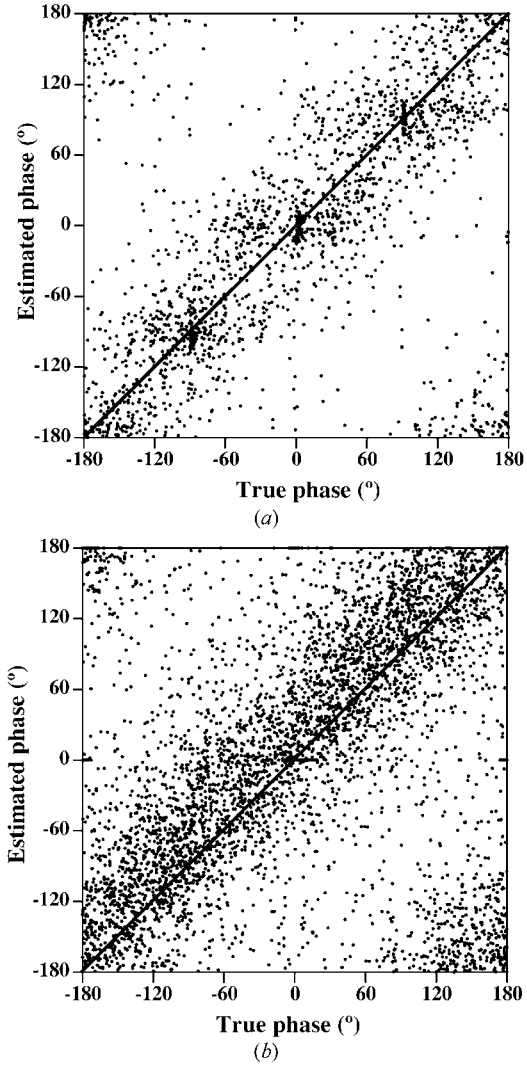

Figure 2
Scatter diagrams of the estimated phase (α) *versus* the true phase (ψ), as well as the line $\psi=\alpha$, for (*a*) the top 2000 reflections from AXE II and (*b*) the top 5000 reflections from AdoHcy having the largest values of Q calculated by formulas (38) and (39).

$$\bar{Q}\cos\bar{\alpha}=W_2R\bar{S}\cos(\varphi-\lambda)+2T_1T_3\cos(\varphi+\lambda+\mu), \quad (40)$$

$$\bar{Q}\sin\bar{\alpha}=-W_2R\bar{S}\sin(\varphi-\lambda)-2T_1T_3\sin(\varphi+\lambda+\mu), \quad (41)$$

where W_1 and W_2 are defined by (21) and (22), and T_1, T_2 and T_3 are calculated *via* the cosine law (see Fig. 1):

$$T_1=\frac{|F_D^+|^2+|F_S+F_S''|^2-|F_P|^2}{2|F_D^+||F_S+F_S''|}, \quad (42)$$

$$T_2=\frac{|F_D^-|^2+|F_S-F_S''|^2-|F_P|^2}{2|F_D^-||F_S-F_S''|}, \quad (43)$$

$$T_3=\frac{|F_D^+|^2+|F_D^-|^2-4|F_S''|^2}{2|F_D^+||F_D^-|}. \quad (44)$$

Equations (36) and (37) imply: (i) when Q is large, $\psi\approx\alpha$; (ii) when $\bar{Q}$ is large, $\bar{\psi}\approx\bar{\alpha}$.

5. Numerical tests

The probabilistic formulas derived in the previous two sections were tested using experimental SIRAS data from (i) the protein acetylxylan esterase (AXE II) (Ghosh *et al.* 1999), (ii) the protein S-adenosylhomocysteine (AdoHcy) hydrolase (Turner *et al.* 1998). AXE II belongs to space group $P2_12_12_1$, with unit-cell parameters $a=34.95$, $b=61.05$ and $c=72.61$ Å. AdoHcy belongs to space group $C222$, with unit-cell parameters $a=91.93$, $b=168.02$, $c=137.77$ Å. The heavy-atom substructures, consisting of 4 iodine atoms for AXE II and 30 selenium atoms for AdoHcy, respectively, were determined by the direct method *Shake-and-Bake* (Weeks *et al.* 1994) using SAS data alone. The basic structural information for these two proteins is listed in Table 1.

5.1. The probabilistic estimate of doublets

The probabilistic formulas (24), (28) and (32) were tested using the experimental SIRAS data listed in Table 1. The parameters A_m and η_m, $m=1,2,3$, which define the conditional probability distributions of the doublets ω_m, were calculated using formulas (25) and (26), (29) and (30), (33) and (34), respectively. All reflections were arranged in descending order of values of the A_m's, and partitioned into eight equal groups (with the exception of the last group) for both proteins. For each group, the average value, $\langle A_m\rangle$, and the corresponding average value of the doublet error, $\langle|\omega_m-\eta_m|\rangle$, were calculated and are reported in Table 2 for (*a*) AXE II, (*b*) AdoHcy.

Table 2 shows that probabilistic formulas (24), (28) and (32) yield reliable (and unique) estimates of tens of thousands of doublets ω_m for both proteins. There is a strong correlation between average doublet error, $\langle|\omega_m-\eta_m|\rangle$, and its concentration parameter value, $\langle A_m\rangle$, *i.e.* the larger the value of $\langle A_m\rangle$, the smaller the average doublet error $\langle|\omega_m-\eta_m|\rangle$.

5.2. The probabilistic estimate of individual phases

The probabilistic formula (36), together with formulas (38) and (39), was tested with the same proteins listed in Table 1. It should be mentioned that the following steps were taken to

deal with the errors arising from data measurements, anomalous substructure determination and the 'lack of closure': (i) the native data, the anomalous scattering derivative data and the calculated heavy-atom substructure data were put on an absolute scale; (ii) a threshold cutoff $|F_{\rm obs}| > 3\sigma(|F_{\rm obs}|)$ was applied to all reflections; (iii) for those 'lack of closure' reflections, values of T_m, $m = 1, 2, 3$, calculated from formulas (42)–(44) were truncated by means of

$$\begin{aligned} T_m &= 0.8 \quad \text{if} \quad T_m > 1.0, \\ T_m &= -0.8 \quad \text{if} \quad T_m < -1.0. \end{aligned} \tag{45}$$

The parameters Q and α, which define the conditional probability distribution (36) of the phase ψ, were calculated using formulas (38) and (39). All reflections were then arranged in descending order of values of the Q's. The average value of Q, $\langle Q \rangle$, and the corresponding average magnitude of the phase error, $\langle |\psi - \alpha| \rangle$, in eight cumulative groups, are calculated and reported in Table 3. A scatter diagram of the estimated phase (α) *versus* the true phase (ψ), as well as the line $\psi = \alpha$, is plotted in Fig. 2 for (*a*) the top 2000 reflections from AXE II and (*b*) the top 5000 reflections from AdoHcy having the largest values of Q. Table 3 shows again the correlation between the concentration parameter Q and its corresponding individual phase error for both proteins, *i.e.* the larger the concentration parameter, the smaller the individual phase error. Fig. 2 shows that (i) the probability formula yields reliable estimates of a large number of individual phases ranging from -180 to 180°, (ii) the probability estimate is unbiased, and (iii) most of the points plotted are centered within a narrow band around $\psi = \alpha$ or around the upper left corner or the lower right corner.

6. Conclusions

In this paper, the goal of integrating the techniques of direct methods, in particular the appropriate advances in the mathematical formalism, with SIRAS is realized. Specifically, the joint probability distribution of three structure factors is obtained. This joint distribution leads directly to the conditional distributions of the two-phase structure invariants, given selected magnitudes. In particular, the formula for estimating individual phases, when a heavy-atom substructure is known, is obtained. The results of the initial applications, using the experimental SIRAS data from the proteins AXE II and AdoHcy, are promising.

The evidence in the previous section strongly shows that it is possible to provide good initial phases for unknown macromolecular structures by probabilistic techniques, provided that SIRAS data are available. It should be stressed that we assume that the substructure has been determined from the SAS data by some direct-methods technique. It is worth pointing out that low resolution and incompleteness of data are no longer major problems, since the probabilistic estimates are based on the individual reflections only. This implies that our method can be applied to very low resolution data, provided only that the heavy-atom substructure is known.

The authors wish to express their appreciation to Dr Blessing for valuable discussions, and to Drs Ghosh and Howell for providing them with test data. This research was supported by NIH grant GM-46733.

References

Deacon, A. M., Ni, Y. S., Coleman, W. G. Jr & Ealick, S. E. (2000). *Structure*, **8**, 453–462.
Fan, H.-F., Han, F.-S., Qian, J.-Z. & Yao, J.-X. (1984). *Acta Cryst.* A**40**, 489–495.
Fortier, S., Fraser, M. E. & Moore, N. J. (1986). *Acta Cryst.* A**42**, 149–156.
Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.
Furey, W., Chandrasekhar, K., Dyda, F. & Sax, M. (1990). *Acta Cryst.* A**46**, 560–567.
Ghosh, D., Erman, M., Sawicki, M., Lala, P., Weeks, D. R., Li, N., Pangborn, W., Thiel, D. J., Jörnvall, H., Gutierrez, R. & Eyzaguirre, J. (1999). *Acta Cryst.* D**55**, 779–784.
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. & Siliqi, D. (2001). *Acta Cryst.* A**57**, 700–707.
Hauptman, H. A. (1982*a*). *Acta Cryst.* A**38**, 289–294.
Hauptman, H. A. (1982*b*). *Acta Cryst.* A**38**, 632–641.
Karle, J. & Hauptman, H. A. (1958). *Acta Cryst.* **11**, 264–269.
Kartha, G. (1975). In *Anomalous Scattering*, edited by S. Ramaseshan & S. C. Abrahams. Copenhagen: Munksgaard.
Turner, M. A., Yuan, C. S., Borchardt, R. T., Hershfield, M. S., Smith, G. D. & Howell, P. L. (1998). *Nature Struct. Biol.* **5**, 369–376.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.