KEULEN, E. (1969). Doctoral thesis, Groningen Univ., The Netherlands.
MATHIESON, A. McL. (1968). *Rev. Sci. Instrum.* **39**, 1834–1837.
MATHIESON, A. McL. (1979). *Acta Cryst.* **A35**, 50–57.
MATHIESON, A. McL. (1982). *Acta Cryst.* **A38**, 378–387.
MATHIESON, A. McL. (1984). *J. Appl. Cryst.* **17**, 207–209.
MATHIESON, A. McL. (1985). *Acta Cryst.* **A41**, 309–316.
MATHIESON, A. McL. & STEVENSON, A. W. (1984). *Aust. J. Phys.* **37**, 657–665.
MATHIESON, A. McL. & STEVENSON, A. W. (1986*a*). *Acta Cryst.* **A42**, 223–230.
MATHIESON, A. McL. & STEVENSON, A. W. (1986*b*). *Acta Cryst.* **A42**, 435–441.
MATHIESON, A. McL. & STEVENSON, A. W. (1993). *Acta Cryst.* **A49**, 655–661.
ROBINSON, B. W. (1933). *Proc. R. Soc. London Ser. A*, **142**, 422–447.
ROBINSON, B. W. (1934). *Proc. R. Soc. London Ser. A*, **147**, 467–478.
SCHNEIDER, J. R. (1977). *Acta Cryst.* **A33**, 235–243.
STEVENSON, A. W. (1989). *Acta Cryst.* **A45**, 75–85.
STEVENSON, A. W. & PAIN, G. N. (1990). *Aust. J. Phys.* **43**, 793–799.
STOKES, A. R. (1948). *Proc. Phys. Soc. London*, **A61**, 382–391.

Acta Cryst. (1995). **A51**, 398–404

The *Ab Initio* Crystal Structure Solution of Proteins by Direct Methods. IV. The Use of the Partial Structure

BY CARMELO GIACOVAZZO

Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 *Bari, Italy*

AND JAVIER GONZALEZ PLATAS

Departamento de Fisica Fundamental y Experimental, Universidad de La Laguna, E-38203 *La Laguna, Tenerife, Spain*

(*Received* 9 *May* 1994; *accepted* 5 *December* 1994)

Abstract

A probabilistic formula originally designed for small molecules, which allows the recovery of the complete from a partial structure [Giacovazzo (1983). *Acta Cryst.* **A39**, 685–692], is reconsidered. Experimental tests show that the formula is potentially able to estimate phases accurately provided 30–40% of the electron density is correctly located. The formula may be used for refining the phase values obtained by isomorphous derivative techniques as well as for extending the phasing process to a resolution higher than the derivative resolution.

Symbols and notation

Papers by Giacovazzo, Siliqi & Ralph (1994), Giacovazzo, Siliqi & Spagna (1994) and Giacovazzo, Siliqi & Zanotti (1995) are referred to as papers I, II and III, respectively.

The symbols and the notation are basically those described in paper III. Additional symbols are necessary here and they are listed below:

$F_{\pi,\mathbf{h}} = |F_{\pi,\mathbf{h}}| \exp(i\varphi_{\pi,\mathbf{h}})$ Structure factor of a partial structure. The subscript p is used in papers I–III, as well as in this paper, to indicated protein.

$[\sigma_2^3/\sigma_3^2]_\pi$ (Statistically equivalent) number of atoms of the partial structure for the primitive unit cell.

$[\sigma_2^3/\sigma_3^2]_q$ (Statistically equivalent) number of atoms of the difference structure obtained by subtracting the partial from the protein structure.

$E''_{\mathbf{h}}$ Structure factors of the protein structure pseudo-normalized with respect to the difference structure.

$E''_{\pi,\mathbf{h}}$ Structure factors of the partial structure pseudo-normalized with respect to the difference structure.

Introduction

According to the tangent formula (Karle & Hauptman, 1956),

$$\tan\theta_{\mathbf{h}} = \frac{\sum_{j=1}^{r} G_j \sin(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j})}{\sum_{j=1}^{r} G_j \cos(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j})} = \frac{T_{\mathbf{h}}}{B_{\mathbf{h}}}. \qquad (1)$$

$\theta_{\mathbf{h}}$ is the most probable value of $\varphi_{\mathbf{h}}$. Its reliability depends on the concentration parameter

$$\alpha_{\mathbf{h}} = (T_{\mathbf{h}}^2 + B_{\mathbf{h}}^2)^{1/2}. \qquad (2)$$

Relationship (1) has practically solved the phase problem for small molecules. Its application to two small proteins

Acta Crystallographica Section A
ISSN 0108-7673 © 1995

(APP, a 36-residue hormone, and rubredoxin from *Desulfovibrio vulgaris*), both with data up to atomic resolution, attained notable success. Relation (1) is strictly connected to the Sayre–Hughes equation (Sayre, 1952; Hughes, 1953):

$$E_{\mathbf{h}} = L^{-1}\sum_{\mathbf{k}} E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}, \qquad (3)$$

which, with respect to (1), imposes additional restraints on the moduli of the structure factors.

Specific reasons make it difficult to apply (1) to proteins of usual size: (*a*) the flatness of the probability distribution $P(\Phi)$; (*b*) the limited data resolution; and (*c*) the difficulty in finding the correct phase set, if obtained, among the various trial solutions.

The problem has been reconsidered by Giacovazzo, Guagliardi, Ravelli & Siliqi (1994). Their results are summarized here:

(*a*) In the absence of phase information,

$$z_{\mathbf{h}} = \langle\alpha_{\mathbf{h}}\rangle/\sigma_{\alpha_{\mathbf{h}}} \qquad (4)$$

may be considered as the signal-to-noise ratio. $\langle\alpha_{\mathbf{h}}\rangle$ is the expected value of $\alpha_{\mathbf{h}}$ given by

$$\langle\alpha_{\mathbf{h}}\rangle = \sum_{j=1}^{r} G_jD_1(G_j) \qquad (5)$$

and $\sigma^2_{\alpha_{\mathbf{h}}}$ (Cascarano, Giacovazzo, Burla, Nunzi & Polidori, 1984) is the variance of $\alpha_{\mathbf{h}}$ given by

$$\sigma^2_{\alpha_{\mathbf{h}}} = \tfrac{1}{2}\sum_{j=1}^{r} G_j^2[1 + D_2(G_j) - 2D_1^2(G_j)]. \qquad (6)$$

(*b*) The statistical solvability criterion was formulated, according to which (1) can be successfully applied to a given set of diffraction data if the relation

$$z \geq T_r \qquad (7)$$

is satisfied by a sufficiently high percentage of large normalized structure factors. T_r represents an acceptable lower limit for the signal-to-noise ratio (say $T_r \simeq 3$).

(*c*) For proteins of usual size,

$$z \leq T_r$$

for a large percentage of reflections. Under these conditions, the Sayre–Hughes equation is practically violated: the correct set of phases is not obtained by application of the tangent formula and the correct solution cannot be recognized among the others.

In paper I, attention was focused on the case in which diffraction data of one isomorphous derivative are also available. In this case, a mathematical procedure (Hauptman, 1982) can be used that integrates direct methods and isomorphous-replacement techniques. Then the triplet reliability parameter G is replaced by (Giacovazzo, Cascarano & Zheng, 1988)

$$A = 2[\sigma_3/\sigma_2^{3/2}]_p R_1R_2R_3 + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_1'\Delta_2'\Delta_3'.$$

The α parameter is accordingly modified. When A is used, condition (7) is satisfied by a sufficiently high percentage of large normalized structure factors. This suggested that the *ab initio* direct solution of proteins is feasible when diffraction data from one isomorphous derivative are additionally available.

Papers II and III were devoted to identifying a practical procedure for the *ab initio* phasing of proteins. It was shown that the phasing process can be extended to about 40% of the measured reflections (up to the derivative resolution) and this can provide interpretable electron-density maps. Since the procedure is highly automatic, thousands of phases can be available in a limited computing time.

In spite of this remarkable success, some drawbacks still limit the applicability of the process:

(1) The quality of the final phases depends on the quality (*i.e.* the degree of isomorphism) of the heavy-atom derivative.

(2) Even if the number of phased reflections is sufficiently large for several practical purposes, a non-negligible number of reflections with $|\Delta| \simeq 0$ but large R values remain unphased. Their contribution to the electron-density map is therefore lost.

(3) The overall phase error is moderately large: its reduction should provide a better definition of the protein envelope.

(4) No method is suggested for extending phases beyond the derivative resolution.

(5) Pseudo-centrosymmetrical phases are obtained in specific space groups.

The aim of this paper is to check the feasibility of a phasing method that exploits as prior information the electron-density map eventually available by application of the procedure described in papers II and III. It will be shown that the method is potentially able to reduce drawbacks (1)–(4).

Direct methods for high-resolution phase refinement

While the *ab initio* solution of protein structures is not within the capacity of traditional direct methods, their efficiency for phase refinement and extension is still under discussion. From the first trials by Reeke & Lipscomb (1969), Weinzierl, Eisenberg & Dickerson (1969) and Coulter (1971), it was clear that a characteristic feature of the tangent formula is the following: a possible moderate improvement of the phases is frequently followed, after a few cycles of refinement, by their deterioration. Phases diverge to a self-consistent incorrect set. The application of the Sayre equation proved more stable even if much more time consuming: therefore, some programs (*e.g. SAYTAN*; Woolfson, 1993) introduce Sayre's-formula restraints in the tangent-formula framework. A more general approach has been followed by Main (1990): the electron-density map is improved by combining information from real and reciprocal spaces. The solution of large non-linear systems, as required by the Sayre equation, is circumvented by the use of the conjugate-gradient method to calculate shifts of the electron-density map. The information so obtained is combined (Cowtan & Main, 1993) with solvent-flattening techniques (Wang, 1985), histogram matching (see Lunin, 1993), non-crystallographic symmetry averaging (Bricogne, 1974) and the use of a partial structure according to the method of Read (1986). The applications of such a method to practical cases show that the improvement of the electron-density map is a product of the simultaneous use of the different techniques.

A different point of view may be introduced. Let us suppose that phase estimates (for example, by isomorphous-derivative techniques) are available for a subset of reflections and that the calculated electron-density map is able to reveal the main features of the structure. The map may be supposed not to be interpretable in terms of chain tracing but showing the general envelope of the molecule. This envelope may be considered as the prior information for the subsequent steps: in particular, its inverse Fourier transform may be calculated and the values F_π are derived for the various structure factors. Then triplet invariants can be estimated *via* distributions like

$$P\big(\varphi_{p,\mathbf{h}}, \varphi_{p,\mathbf{k}}, \varphi_{p,\mathbf{h-k}} \big| |F_{p,\mathbf{h}}|, |F_{p,\mathbf{k}}|, |F_{p,\mathbf{h-k}}|, |F_{\pi,\mathbf{h}}|, |F_{\pi,\mathbf{k}}|, |F_{\pi,\mathbf{h-k}}|, \varphi_{\pi,\mathbf{h}}, \varphi_{\pi,\mathbf{k}}, \varphi_{\pi,\mathbf{h-k}}\big), \tag{8}$$

derived by Giacovazzo (1983), rather than *via*

$$P\big(\varphi_{p,\mathbf{h}}, \varphi_{p,\mathbf{k}}, \varphi_{p,\mathbf{h-k}} \big| |F_{p,\mathbf{h}}|, |F_{p,\mathbf{k}}|, |F_{p,\mathbf{h-k}}|\big) \tag{9}$$

used by tangent formula (1). The advantage of (8) with respect to the other methods may be summarized as:

(*a*) The electron-density map is divided into two regions; the first coincides with the assumed partial structure, the second is 'flattened' to zero and gives vanishing contribution to the values of F_π.

(*b*) The distribution can take full advantage of the known partial structure, which on the contrary is neglected in (9).

(*c*) The prior information proved to lead to highly accurate estimates of the phases (Camalli, Giacovazzo & Spagna, 1985; Burla, Cascarano, Fares, Giacovazzo, Polidori & Spagna, 1989), at least for small-molecule structures.

The problem is now: is the supplementary information provided by (8) sufficient for reliably extending and refining phases of macromolecules? The answer is not easy: the effectiveness of the process depends on the accuracy of the starting phases φ_π, and therefore on the general correctness of the envelope, and on the complexity of the entire structure. In this paper, we want to explore first the feasibility of the method by working in an ideal and therefore well controlled situation.

The probabilistic formula to apply, derived from (8), is

$$E''_{\mathbf{h}} \simeq E''_{\pi,\mathbf{h}} + [\sigma_3/\sigma_2^{3/2}]_q \sum_{\mathbf{k}} (E''_{\mathbf{k}} - E''_{\pi,\mathbf{k}})(E''_{\mathbf{h-k}} - E''_{\pi,\mathbf{h-k}}), \tag{10}$$

which, in terms of phases, is equivalent to

$$\tan\theta_{\mathbf{h}} = T_\pi/B_\pi, \tag{11}$$

where

$$\begin{aligned} T_\pi = {} & 2R''_{\mathbf{h}}\{R''_{\pi,\mathbf{h}} \sin\varphi_{\pi,\mathbf{h}} + [\sigma_3/\sigma_2^{3/2}]_q \\ & \times \sum_{\mathbf{k}} [R''_{\mathbf{k}} R''_{\mathbf{h-k}} \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h-k}}) \\ & - R''_{\pi,\mathbf{k}} R''_{\mathbf{h-k}} \sin(\varphi_{\pi,\mathbf{k}} + \varphi_{\mathbf{h-k}}) \\ & - R''_{\mathbf{k}} R''_{\pi,\mathbf{h-k}} \sin(\varphi_{\mathbf{k}} + \varphi_{\pi,\mathbf{h-k}}) \\ & + R''_{\pi,\mathbf{k}} R''_{\pi,\mathbf{h-k}} \sin(\varphi_{\pi,\mathbf{k}} + \varphi_{\pi,\mathbf{h-k}})]\} \\ B_\pi = {} & 2R''_{\mathbf{h}}\{R''_{\pi,\mathbf{h}} \cos\varphi_{\pi,\mathbf{h}} + [\sigma_3/\sigma_2^{3/2}]_q \\ & \times \sum_{\mathbf{k}} [R''_{\mathbf{k}} R''_{\mathbf{h-k}} \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h-k}}) + \ldots]\}. \end{aligned}$$

$\theta_{\mathbf{h}}$ is the most probable value of $\varphi_{\mathbf{h}}$ and

$$\alpha_{\pi,\mathbf{h}} = (T_\pi^2 + B_\pi^2)^{1/2} \tag{12}$$

is its reliability parameter.

The statistical *z* test when a partial structure is available

In order to estimate the efficiency of (10), we should calculate, in accordance with (4),

$$z_{\pi,\mathbf{h}} = \langle \alpha_{\pi,\mathbf{h}} \rangle / \sigma_{\alpha_{\pi,\mathbf{h}}}. \tag{13}$$

$\langle \alpha_{\pi,\mathbf{h}} \rangle$ and $\sigma_{\alpha_{\pi,\mathbf{h}}}$ may be derived according to the following procedure:

(1) First we derive, from equation (19) of the paper by Giacovazzo (1983), the marginal distribution $P(\varphi_{\mathbf{k}}, \varphi_{\mathbf{h-k}} | R''_{\mathbf{h}}, R''_{\mathbf{k}}, R''_{\mathbf{h-k}}, R''_{\pi,\mathbf{h}}, R''_{\pi,\mathbf{k}}, R''_{\pi,\mathbf{h-k}}, \varphi_{\pi,\mathbf{h}}, \varphi_{\pi,\mathbf{k}}, \varphi_{\pi,\mathbf{h-k}})$, which is obtained by integrating (8) over $\varphi_{\mathbf{h}}$. Indeed, in our case, the phase $\varphi_{\mathbf{h}}$ is supposed to be unknown, in accordance with the fact that we are interested in $\langle \alpha_{\pi,\mathbf{h}} \rangle$. For the sake of simplicity, we neglect terms of order $[\sigma_3/\sigma_2^{3/2}]_q$ and we find that $(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h-k}})$ is distributed according to the Von Mises distribution

$$M(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h-k}}; \varphi_{\pi,\mathbf{k}} + \varphi_{\pi,\mathbf{h-k}}, q_{1,\mathbf{k}}),$$

where $q_{1,\mathbf{k}}$ satisfies the relation

$$D_1(q_{1,\mathbf{k}}) = D_1(2R''_{\mathbf{k}} R''_{\pi,\mathbf{k}}) D_1(2R''_{\mathbf{h-k}} R''_{\pi,\mathbf{h-k}}).$$

$(\varphi_{\pi,\mathbf{k}} + \varphi_{\pi,\mathbf{h-k}})$ is the expected value of $(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h-k}})$ and $q_{1,\mathbf{k}}$ is the concentration parameter of the distribution.

(2) In an analogous way, we obtain that $(\varphi_{\pi,\mathbf{k}} + \varphi_{\mathbf{h-k}})$ is distributed according to the Von Mises distribution

$$M(\varphi_{\pi,\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}};\, \varphi_{\pi,\mathbf{k}} + \varphi_{\pi,\mathbf{h}-\mathbf{k}},\, q_{2,\mathbf{k}}),$$

where $q_{2,\mathbf{k}} = 2R''_{\mathbf{h}-\mathbf{k}}R''_{\pi,\mathbf{h}-\mathbf{k}}$.

(3) Also, $(\varphi_{\mathbf{k}} + \varphi_{\pi,\mathbf{h}-\mathbf{k}})$ is distributed according to

$$M(\varphi_{\mathbf{k}} + \varphi_{\pi,\mathbf{h}-\mathbf{k}};\, \varphi_{\pi,\mathbf{k}} + \varphi_{\pi,\mathbf{h}-\mathbf{k}},\, q_{3,\mathbf{k}}),$$

where $q_{3,\mathbf{k}} = 2R''_{\mathbf{k}}R''_{\pi,\mathbf{k}}$.

(4) We recall that the distribution of the modulus α of the resultant of r complex vectors $Q_j\exp(i\nu_j)$ under the hypothesis that Q_j are distributed according to the Von Mises distribution $M(\nu_j;\, \theta, q_j)$ is the normal distribution (Cascarano, Giacovazzo & Guagliardi, 1992) $N(\alpha;\, \langle\alpha\rangle, \sigma^2)$, where

$$\langle\alpha\rangle = \sum_{j=1}^{r} Q_j D_1(q_j)$$

$$\sigma^2 = \tfrac{1}{2}\sum_{j=1}^{r} Q_j^2[1 + D_2(q_j) - 2D_1^2(q_j)]. \qquad (14)$$

(5) We apply the above result to the sets of vectors

$$\begin{aligned}&\sum_{\mathbf{k}} Q_{1,\mathbf{k}}\exp[i(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})],\\ &\sum_{\mathbf{k}} Q_{2,\mathbf{k}}\exp[i(\varphi_{\pi,\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})],\\ &\sum_{\mathbf{k}} Q_{3,\mathbf{k}}\exp[i(\varphi_{\mathbf{k}} + \varphi_{\pi,\mathbf{h}-\mathbf{k}})],\end{aligned} \qquad (15)$$

where

$$\begin{aligned}Q_{1,\mathbf{k}} &= 2[\sigma_3/\sigma_2^{3/2}]_q R''_{\mathbf{k}}R''_{\mathbf{h}-\mathbf{k}},\\ Q_{2,\mathbf{k}} &= 2[\sigma_3/\sigma_2^{3/2}]_q R''_{\pi,\mathbf{k}}R''_{\mathbf{h}-\mathbf{k}},\\ Q_{3,\mathbf{k}} &= 2[\sigma_3/\sigma_2^{3/2}]_q R''_{\mathbf{k}}R''_{\pi,\mathbf{h}-\mathbf{k}}.\end{aligned}$$

Then,

$$\begin{aligned}\langle\alpha_{,\mathbf{h}}\rangle = 2R''_{\pi,\mathbf{h}} + \sum_{\mathbf{k}}[&Q_{1,\mathbf{k}}D_1(q_{1,\mathbf{k}})\\ &- Q_{2,\mathbf{k}}D_1(q_{2,\mathbf{k}}) - Q_{3,\mathbf{k}}D_1(q_{3,\mathbf{k}}) + Q_{4,\mathbf{k}}],\end{aligned} \qquad (16)$$

where

$$Q_{4,\mathbf{k}} = 2[\sigma_3/\sigma_2^{3/2}]_q R''_{\pi,\mathbf{k}}R''_{\pi,\mathbf{h}-\mathbf{k}}.$$

$\langle\alpha_{\pi,\mathbf{h}}\rangle$ reduces to $\langle\alpha_{\mathbf{h}}\rangle$ when no partial structure is available.

(6) The value of $\sigma^2_{\alpha_{\pi,\mathbf{h}}}$ may be derived by applying (14) in turn to the terms in (15) and then summing the contributions.

Table 1. *Code name, space group and crystallochemical data for test structures*

Structure code	Space group	Molecular formula	Z
APP[(a)]	$C2$	$C_{190}N_{53}O_{58}Zn$	4
BPTI[(b)]	$P2_12_12_1$	$S_9O_{149}N_{84}C_{289}$	4
LYSO[(c)]	$P4_32_12$	$S_{10}O_{286}N_{193}C_{613}$	8
MYO[(d)]	$P2_1$	$FeS_4O_{389}N_{220}C_{817}$	2
M-FABP[(e)]	$P2_12_12_1$	$C_{667}N_{170}O_{261}S_3$	4
E2[(f)]	$F432$	$C_{1170}N_{310}O_{366}S_7$	96

References: (*a*) Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983); (*b*) data by courtesy of R. Huber, MPI Martinsried, Germany; (*c*) data by courtesy of C. Betzel, ENBL, Hamburg, Germany; (*d*) Hartmann, Steigemann, Reuscher & Parak (1987); (*e*) Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992); (*f*) Mattevi, Obmolova, Schulze, Kalk, Westphal, De Kok & Hol (1992).

Table 2. *Parameters defining protocol for calculations* (*see the main text for the symbols*)

Code	RES (Å)	N	NREFL	NLAR	NTRIP
APP	0.99	413	17058	700	8907
BPTI	1.00	1860	17300	700	21759
LYSO	1.69	7720	13622	700	24899
MYO	1.50	2648	15903	1100	26796
M-FABP	2.14	4076	7769	700	23012
E2	3.00	40783	8136	400	23820

The statistical solvability criterion has been applied to the experimental data of the structures quoted in Table 1. For each test structure, we give in Table 2 the resolution of the diffraction data (RES), the number of atoms (statistically calculated) in the primitive unit cell (N), the number of measured reflections (NREFL), the number of large normalized structure factors (NLAR) among which triplet invariants are calculated, and the total number of triplets that contribute to the various α values (NTRIP). For each structure, we calculated the $z_{\mathbf{h}}$ values corresponding to the NLAR reflections according to definition (4) and the $z_{\pi,\mathbf{h}}$ values according to definition (13).

Two different amounts of prior information were used for $z_{\pi,\mathbf{h}}$ corresponding to different values of the diffraction ratio $DR_\pi = [\sigma_2]_\pi/[\sigma_2]_p = 0.20, 0.40$. In Figs. 1–6, we show the $P(z)$ and $P(z_\pi)$ curves. In general, $P(z)$ curves do not satisfy the statistical solvability criterion: on the contrary, the $P(z_\pi)$ curves are remarkably shifted towards the right and satisfy the criterion. The only exception occurs for APP for which the prior information does not improve the distribution of the signal-to-noise ratio. At the moment, we are unable to explain this discordant result. The shifts relative to the $P(z_\pi)$ curves increase with the amount of prior information: therefore, more information should make (10) more efficient.

In order to collect the above observations in a simple sentence, we can conclude that the Sayre–Hughes relation (3) is expected to be violated for all test structures except APP, while relation (10) is expected to be satisfied for all test structures. In order to check this conclusion, we calculated the following figures of merit:

$$\mathrm{FOM} = \frac{\sum_{\mathbf{h}} |E_{\mathbf{h}} - E_{\mathbf{h}_{\mathrm{cal}}}|}{\sum_{\mathbf{h}} |E_{\mathbf{h}}|} \quad \mathrm{FOM}_\pi = \frac{\sum_{\mathbf{h}} |E''_{\mathbf{h}} - E''_{\mathbf{h}_{\mathrm{cal}}}|}{\sum_{\mathbf{h}} |E''_{\mathbf{h}}|}. \qquad (17)$$

In (17), $E_{\mathbf{h}}$ and $E''_{\mathbf{h}}$ stand for $R_{\mathbf{h}}\exp(i\varphi_{\mathbf{h}})$ and $R''_{\mathbf{h}}\exp(i\varphi_{\mathbf{h}})$, respectively; $\varphi_{\mathbf{h}}$ is the true phase (derived

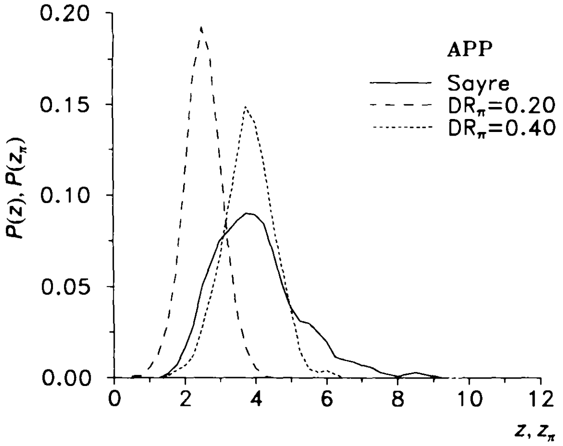

Fig. 1. APP: $P(z)$ and $P(z_\pi)$ curves.

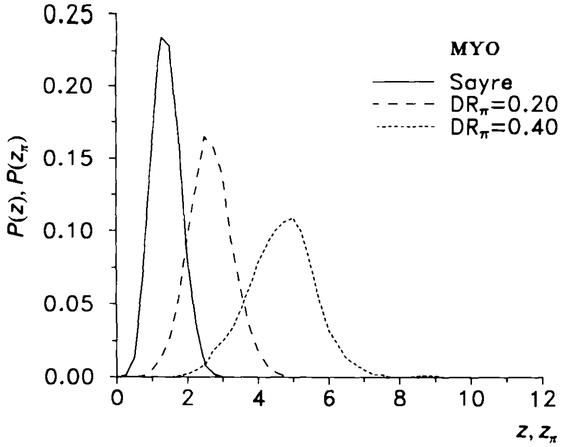

Fig. 4. MYO: $P(z)$ and $P(z_\pi)$ curves.

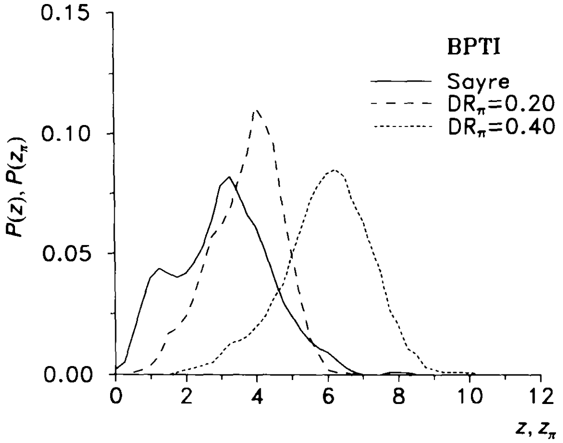

Fig. 2. BPTI: $P(z)$ and $P(z_\pi)$ curves.

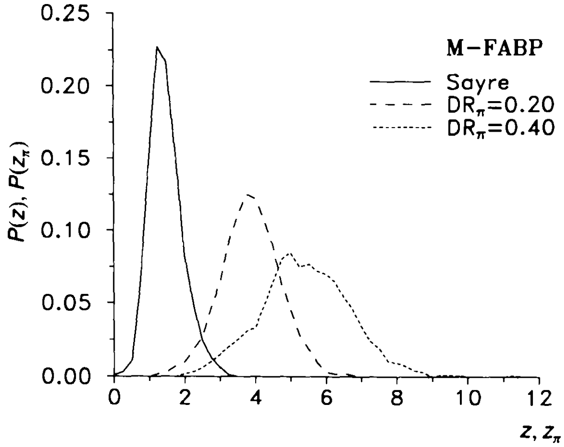

Fig. 5. M-FABP: $P(z)$ and $P(z_\pi)$ curves.

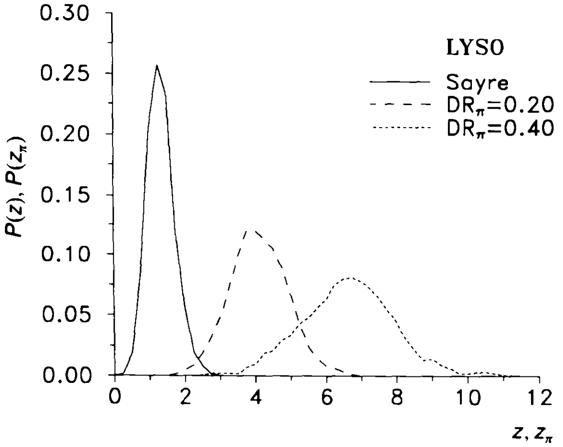

Fig. 3. LYSO: $P(z)$ and $P(z_\pi)$ curves.

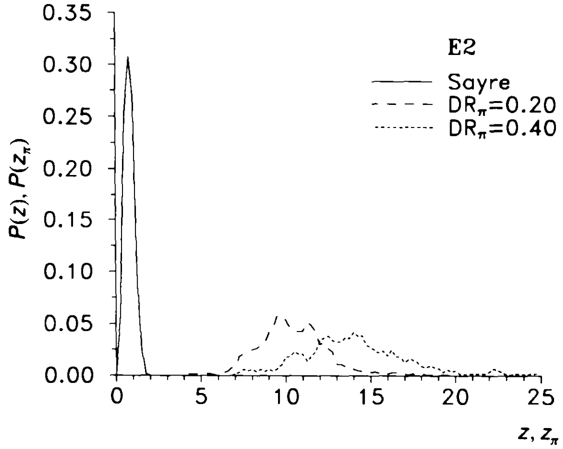

Fig. 6. E2: $P(z)$ and $P(z_\pi)$ curves.

Table 3. FOM *and* FOM_π *values for the test structures*

Code	FOM	FOM_π ($\mathrm{DR}_\pi = 0.20$)	FOM_π ($\mathrm{DR}_\pi = 0.40$)
APP	0.512	0.515	0.490
BPTI	0.783	0.579	0.496
LYSO	0.847	0.696	0.598
MYO	0.800	0.632	0.497
M-FABP	0.863	0.695	0.570
E2	0.910	0.567	0.477

Table 4. $\langle|\Delta\varphi^o|\rangle = \langle|\varphi^o_{\mathrm{calc}} - \varphi^o_{\mathrm{true}}|\rangle$: *mean phase error for Sayre–Hughes relationship*; $\langle|\Delta\varphi^o|\rangle_\pi$: *mean phase error for* (11)

Code	$\langle|\Delta\varphi^o|\rangle$ (°)	$\langle|\Delta\varphi^o|\rangle_\pi$ (°) ($\mathrm{DR}_\pi = 0.20$)	$\langle|\Delta\varphi^o|\rangle_\pi$ (°) ($\mathrm{DR}_\pi = 0.40$)
APP	23.7	20.0	18.4
BPTI	27.1	21.5	16.4
LYSO	45.1	32.1	25.4
MYO	42.6	28.9	20.6
M-FABP	46.3	32.7	23.3
E2	52.8	23.4	17.7

from the refined published crystal structure). *R* and *R″* are respectively normalized and pseudonormalized magnitudes derived from measurements.

When FOM is calculated for the Sayre–Hughes relationship, $E_{\mathbf{h}_{\mathrm{cal}}}$ is obtained from the right-hand side of (3) by using true phases $(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})$. When FOM_π is calculated for (10), $E''_{\mathbf{h}_{\mathrm{cal}}}$ is obtained from the right-hand side of (10) by using the true values $\varphi_{\mathbf{k}}$, $\varphi_{\mathbf{h}-\mathbf{k}}$, $\varphi_{\pi,\mathbf{h}}$, $\varphi_{\pi,\mathbf{k}}$, $\varphi_{\pi,\mathbf{h}-\mathbf{k}}$. Large values of FOM and FOM_π involve remarkable deviations (in terms of moduli and phases) of the calculated *E*'s from the observed ones and therefore indicate violation of (3) and (10). The results are shown in Table 3. FOM values are quite large, thus confirming that the Sayre–Hughes relation is not satisfied. FOM_π's are remarkably smaller than the corresponding FOM's: the situation still improves when DR_π increases. Comparison between the numerical values of FOM and FOM_π confirms the significance of Figs. 1–6: a good correlation can be found between the shifts to the right of the distribution $P(z_\pi)$ and the differences $\mathrm{FOM}-\mathrm{FOM}_\pi$. For example, $\mathrm{FOM}_\pi \simeq \mathrm{FOM}$ for APP.

The values of FOM and FOM_π in Table 3 suggest that the accuracy of the phases determined *via* (11) is expected to be higher than for phases fixed by (3); the average accuracy should increase with DR_π. This expectation is confirmed by a supplementary test (see Table 4): we calculated the average phase error $\langle|\Delta\varphi^o|\rangle = \langle|\varphi^o_{\mathrm{cal}} - \varphi^o_{\mathrm{true}}|\rangle$ corresponding to FOM and FOM_π. The larger the amount of prior information, the better the phase estimates are.

A question could arise: is the value $\varphi_{\mathbf{h}_{\mathrm{cal}}}$, as calculated from (11), closer on average to $\varphi_{\mathbf{h}_{\mathrm{true}}}$ than the value $\varphi_{\pi,\mathbf{h}}$? In Table 5, we show the values $\langle|\varphi^o_{\mathrm{true}} - \varphi^o_\pi|\rangle$ for the different values of DR_π. If these are compared with the average phase errors in Table 4, the important role of the triplet contribution in the phasing process is realized. Thus, phases are estimated through (11) much better than through the Sim (1959) relationship.

Table 5. *Average phase errors according to the Sim (1959) relationship*

Code	$|\varphi^o_{\mathrm{true}} - \varphi^o_\pi|$ ($\mathrm{DR}_\pi = 0.20$)	$|\varphi^o_{\mathrm{true}} - \varphi^o_\pi|$ ($\mathrm{DR}_\pi = 0.40$)
APP	39.5	29.8
BPTI	47.1	24.4
LYSO	46.5	30.1
MYO	41.6	25.5
M-FABP	43.2	29.4
E2	30.7	20.8

Tangent refinement

It is worthwhile noting that in Table 4 $\langle|\Delta\varphi^o|\rangle$ is rather small even for the Sayre relationship. However, it should not be concluded that the Sayre relationship is satisfied. Indeed, the correct criterion for deciding on the violation of (3) or (10) is the inspection of FOM and FOM_π because they simultaneously involve phases and moduli. If this is true, the values of FOM and FOM_π should be useful indicators for foreseeing the behaviour of the tangent procedures. In particular, they should measure the tendency of the tangent formulas (1) and (11) to diverge to self-consistent incorrect sets of phases.

In order to confirm this property, we started phase refinement from correct phase values according to (1) and (11) and we checked the average phase error after convergence was attained. The threshold value TR_α (*i.e.* a reflection is considered phased if $\alpha \geq \mathrm{TR}_\alpha$) is multiplied by 0.65 in each new cycle. The process stops for (3) if $(\sum_{\mathbf{h}} \alpha - \sum_{\mathbf{h}} \alpha_{\mathrm{pc}})/\sum_{\mathbf{h}} \alpha \leq 0.02$, where α_{pc} is the α value in the preceding refinement cycle. When (11) is used, α_π replaces α. The results are shown in Table 6. It may be noted that:

(i) Not all the NLAR reflections are phased at the end of the process. The percentage of phased reflections is small for E2 when the Sayre–Hughes relation is used.

(ii) Phases diverge remarkably, except for LYSO, M-FABP and E2 when $\mathrm{DR}_\pi \neq 0$.

(iii) As a general trend, (11) is more efficient than (1) but is still not satisfactory.

Much better results are obtained by slightly modifying the refinement process. The program stops when $\mathrm{TR}_\alpha \leq 1.5$ (this condition prevents unreliable phase assignments) or when the number of phase reflections is larger than 0.85 x NLAR. This last condition avoids repeated cycles of refinement on the same set of phases: Under these conditions, phases usually move towards autoconsistency and diverge from the true values. The results of the new procedure are shown in Table 7. It may be noted that: (*a*) the number of phased reflections is

Table 6. $\langle|\Delta\varphi^o|\rangle$ *after the application of the tangent formula to true phases*; % *is the percentage of* NLAR *reflections phased by the process*

Code	$\langle\lvert\Delta\varphi^o\rvert\rangle$ (°) (%) Sayre–Hughes relation	$\langle\lvert\Delta\varphi^o\rvert\rangle$ (°) (%) Equation (11) ($\mathrm{DR}_\pi = 0.20$)	$\langle\lvert\Delta\varphi^o\rvert\rangle$ (°) (%) Equation (11) ($\mathrm{DR}_\pi = 0.40$)
APP	43.2	41.7	43.5
	(99)	(99)	(99)
BPTI	79.4	66.4	65.9
	(93)	(98)	(99)
LYSO	40.2	31.2	24.5
	(88)	(91)	(95)
MYO	68.0	62.1	57.6
	(98)	(99)	(100)
M-FABP	41.8	30.0	23.1
	(87)	(89)	(93)
E2	35.2	26.1	19.2
	(53)	(91)	(98)

Table 7. $\langle|\Delta\varphi^o|\rangle$ *after the application of the tangent formula to true phases*; % *is the percentage of* NLAR *reflections phased by the process*; *the minimum threshold for* TR_α *is* 1.5

Code	$\langle\lvert\Delta\varphi^o\rvert\rangle$ (°) (%) Sayre–Hughes relation	$\langle\lvert\Delta\varphi^o\rvert\rangle$ (°) (%) Equation (11) ($\mathrm{DR}_\pi = 0.20$)	$\langle\lvert\Delta\varphi^o\rvert\rangle$ (°) (%) Equation (11) ($\mathrm{DR}_\pi = 0.40$)
APP	27.3	25.9	25.6
	(85)	(88)	(90)
BPTI	36.4	24.5	14.6
	(88)	(92)	(88)
LYSO	32.9	27.5	22.9
	(72)	(86)	(93)
MYO	40.7	25.5	17.4
	(85)	(88)	(92)
M-FABP	33.4	26.1	19.5
	(68)	(84)	(85)
E2	30.8	23.3	16.9
	(32)	(87)	(93)

larger when relation (11) is used; in particular, the Sayre equation is still unable to fix for E2 the phases of about $0.68 \times \mathrm{NLAR}$ reflections; (*b*) the phase error is remarkably smaller for relation (11): the error decreases when DR_π increases.

The above conclusions confirm that the suggestions we derived from Figs. 1–6 are sound: the prior information on part of a crystal structure allows the successful application of (10) to macromolecules; that is, the complete crystal structure may in principle be recovered when a partial structure is available.

Concluding remarks

It has been shown that relationship (10) is potentially able to estimate accurately the phases of a relevant number of reflections provided some prior information is available on part of the structure. As a rule of thumb, prior information on about 30–40% of the structure should make (10) highly efficient.

Relationship (10) can be used in two different ways: (*a*) Combined with the probabilistic techniques described in papers I–III to improve the phasing process for reflections up to isomorphous resolution. In this case, the partial structure constitutes a supplementary derivative, the quality of which depends on the accuracy with which the partial is defined. Tangent refinement of this second derivative will produce phases that may be usefully combined with MIR phases. (*b*) As a stand-alone technique that is particularly useful at resolution higher than the derivative resolution. In both cases, the use of (10) should be cyclic: the initial prior information is used for phase extension and refinement, which, in turn, should provide a better electron-density map and therefore a better partial structure to use as new prior information. The practical results of this procedure will be described in a following paper.

References

Bricogne, G. (1974). *Acta Cryst.* **A30**, 395–405.

Burla, M. C., Cascarano, G., Fares, V., Giacovazzo, C., Polidori, G. & Spagna, R. (1989). *Acta Cryst.* **A45**, 781–786.

Camalli, M., Giacovazzo, C. & Spagna, R. (1985). *Acta Cryst.* **A41**, 605–613.

Cascarano, G., Giacovazzo, C., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* **A40**, 389–394.

Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1992). *Acta Cryst.* **A48**, 859–865.

Coulter. C. L. (1971). *Acta Cryst.* **B27**, 1730–1740.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* **D49**, 148–157.

Giacovazzo, C. (1983). *Acta Cryst.* **A39**, 685–692.

Giacovazzo, C., Cascarano, G. & Zheng, C. (1988). *Acta Cryst.* **A44**, 45–51.

Giacovazzo, C., Guagliardi, A., Ravelli, R. & Siliqi, D. (1994). *Z. Kristallogr.* **209**, 136–142.

Giacovazzo, C., Siliqi, D. & Ralph, A. (1994). *Acta Cryst.* **A50**, 503–505.

Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst.* **A50**, 609–621.

Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *Acta Cryst.* **A51**, 177–188.

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

Hartmann, H., Steigemann, W., Reuscher, H. & Parak, F. (1987). *Eur. Biophys. J.* **14**, 337–348.

Hauptman, H. (1982). *Acta Cryst.* **A38**, 289–294.

Hughes, E. W. (1953). *Acta Cryst.* **6**, 871.

Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.

Lunin, V. Yu. (1993). *Acta Cryst.* **D49**, 90–99.

Main, P. (1990). *Acta Cryst.* **A46**, 372–377.

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

Read, R. J. (1986). *Acta Cryst.* **A42**, 140–149.

Reeke, G. N. & Lipscomb, W. N. (1969). *Acta Cryst.* **B25**, 2614–2623.

Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.

Sim, G. (1959). *Acta Cryst.* **12**, 813–815.

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

Weinzierl, J. E., Eisenberg, D. & Dickerson, R. E. (1969). *Acta Cryst.* **B25**, 380–387.

Woolfson, M. M. (1993). *Acta Cryst.* **D49**, 13–17.

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