gratefully acknowledge the computer time provided by the following institutions: HLRZ Jülich GmbH (Cray Y-MP), Verbund der Hessischen Höchstleistungsrechner (SNI 100) and Rechenzentrum der Universität Marburg (Convex C-230).

References

BARTELL, L. S. & GAVIN, R. M. JR (1964). *J. Am. Chem. Soc.* **86**, 3493–3498.
BENESCH, R. & SMITH, V. H. JR (1970). *Acta Cryst.* A**26**, 586–594.
BONHAM, R. A. & GORUGANTHU, R. R. (1982). *Phys. Rev. A*, **26**, 1–11.
BREITENSTEIN, M., MEYER, H. & SCHWEIG, A. (1985). *Chem. Phys. Lett.* **119**, 120–127.
BUNGE, C. F. (1976). *At. Data Nucl. Data Tables*, **18**, 293–304.
CHAKRAVORTY, S. J., GWALTNEY, S. R., DAVIDSON, E. R., PARPIA, F. A. & FROESE-FISCHER, C. (1993). *Phys. Rev. A*, **47**, 3649–3670.
DOYLE, P. A. & TURNER, P. S. (1968). *Acta Cryst.* A**24**, 390–397.
ESQUIVEL, R. O. & BUNGE, A. V. (1987). *Int. J. Quant. Chem.* **32**, 295–312.
FELLER, D. & DAVIDSON, E. R. (1988). *J. Chem. Phys.* **88**, 7580–7587.
FELLER, D. & DAVIDSON, E. R. (1989). *J. Chem. Phys.* **89**, 1024–1030.
FROESE-FISCHER, C. (1977). *The Hartree–Fock Method for Atoms*. New York: Wiley.
MASLEN, E. N., FOX, A. G. & O'KEEFE, M. A. (1992). *International Tables for Crystallography*, Vol. C, edited by A. J. C. WILSON, pp. 476–511. Dordrecht: Kluwer Academic Publishers.
MEYER, H., MÜLLER, T. & SCHWEIG, A. (1995). *Chem. Phys.* In the press.
NAON, M. & CORNILLE, M. (1973). *J. Phys. B*, **6**, 1347–1356.
PEIXOTO, E. M. A., BUNGE, C. F. & BONHAM, R. A. (1969). *Phys. Rev.* **181**, 322–328.
SASAKI, F. & YOSHIMINE, M. (1974). *Phys. Rev. A*, **9**, 17–25.
SCHMIDER, H., ESQUIVEL, R. O., SAGAR, R. P. & SMITH, V. H. JR (1993). *J. Phys. B*, **26**, 2943–2955.
SHEPARD, R., LISCHKA, H., SZALAY, P. G., KOVAR, T. & ERNZERHOF, M. (1992). *J. Chem. Phys.* **96**, 2085–2098.
SHEPARD, R., SHAVITT, I., PITZER, R. M., COMEAU, D. C., PEPPER, M., LISCHKA, H., SZALAY, P. G., AHLRICHS, R., BROWN, F. B. & ZHAO, J.-G. (1988). *Int. J. Quantum Chem.* S**22**, 149–165.
SIMAS, A. M., SAGAR, R. P., KU, A. C. T. & SMITH, V. H. JR (1988). *Can. J. Chem.* **66**, 1923–1930.
TANAKA, K. & SASAKI, F. (1971). *Int. J. Quantum Chem.* **5**, 157–175.
TAVARD, C. (1965). *Cah. Phys.* **20**, 397–495.
TAVARD, C., NICOLAS, D. & ROUAULT, M. (1967). *J. Chim. Phys.* **64**, 540–554.
THAKKAR, A. J. & SMITH, V. H. JR (1992). *Acta Cryst.* A**48**, 70–71.
WALLER, I. & HARTREE, R. D. (1929). *Proc. R. Soc. London Ser. A*, **124**, 119–142.
WANG, J., SAGAR, R. P., SCHMIDER, H. & SMITH, V. H. JR (1993). *At. Data Nucl. Data Tables*, **53**, 233–269.

Acta Cryst. (1995). A**51**, 177–188

The *Ab Initio* Crystal Structure Solution of Proteins by Direct Methods. III. The Phase Extension Process

BY CARMELO GIACOVAZZO AND DRITAN SILIQI*

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

AND GIUSEPPE ZANOTTI

Dipartimento di Chimica Organica, Università di Padova, Via Marzolo 1, *Padova, Italy*

(*Received* 18 *April* 1994; *accepted* 8 *September* 1994)

Abstract

In two preceding papers [Giacovazzo, Siliqi & Ralph (1994). *Acta Cryst.* A**50**, 503–510; Giacovazzo, Siliqi & Spagna (1994). *Acta Cryst.* A**50**, 609–621], a direct-phasing process was described which proved to be potentially able to solve *ab initio* crystal structures of proteins. The method uses the diffraction data of the native and of one isomorphous derivative. The main limitation of the approach was the small number of phased reflections rather than the quality of the assigned phases. In this paper, it is shown that the phasing process can be extended to about 40% of the measured reflections (up to the derivative resolution) without reducing significantly the quality of the new phases. Of the four test proteins examined, in one case it was possible to obtain fully interpretable electron-density maps.

Symbols and abbreviations

Symbols and notation are basically the same as in papers I and II (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994). Since new symbols are necessary here and for the reader's convenience, we give a combined list below.

$F_p = \|F_p\| \exp(i\varphi)$	Structure factor of the protein
$F_d = \|F_d\| \exp(i\psi)$	Structure factor of the isomorphous derivative
$F_H = F_d - F_p$	Structure factor of the heavy-atom structure (*i.e.* the atoms added to the native protein)

* Permanent address: Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana University, Tirana, Albania.

$\Phi=\varphi_{\mathbf{h}}-\varphi_{\mathbf{k}}-\varphi_{\mathbf{h}-\mathbf{k}}$	
$E_p=R\exp(i\varphi)$	Normalized structure factor of the protein
$E_d=S\exp(i\psi)$	Normalized structure factor of the isomorphous derivative
N	Number of non-H atoms in the primitive cell for the native protein
$\sigma_i=\sum_{j=1}^{N}Z_j^i$	Z_j = atomic number of jth atom
$N_{\rm eq}=\sigma_2^3/\sigma_3^2$	(Statistically equivalent) number of atoms in the primitive unit cell
$[\sigma_2^3/\sigma_3^2]_p$	Value of $N_{\rm eq}$ for the native protein
$[\sigma_2^3/\sigma_3^2]_H$	Value of $N_{\rm eq}$ for the heavy-atom structure
$G=2[\sigma_3/\sigma_2{}^{3/2}]_p\lvert R_{\mathbf{h}}R_{\mathbf{k}}R_{\mathbf{h}-\mathbf{k}}\rvert$	
f_j	Atomic scattering factor of the jth atom
$\Sigma_p=\Sigma_p f_j^2$	The sum is extended to the native protein atoms
$\Sigma_H=\Sigma_H f_j^2$	The sum is extended to the heavy-atom structure
$\Sigma_d=\Sigma_d f_j^2$	The sum is extended to the derivative atoms
$D_i(x)=I_i(x)/I_0(x)$	I_i = modified Bessel function of order i
$E'_d=F_d/\Sigma_H^{1/2}=S'\exp(i\psi)$	Derivative pseudonormalized structure factor
$E'_p=F_p/\Sigma_H^{1/2}=R'\exp(i\varphi)$	Native protein pseudonormalized structure factor
$\Delta=S'-R'$	$\Delta'=S'T-R'$
$T=D_1(2R'S')$	$\sigma=[\sigma_2]_H/[\sigma_2]_p$
CARP	Carp muscle calcium-binding protein
E2	Catalytic domain of *Azotobacter vinlandii* dihydrolipoyl transacetylase
M-FABP	Recombinant human muscle fatty-acid-binding protein

Introduction

In paper I of this series, the *statistical solvability criterion* (Giacovazzo, Guagliardi, Ravelli & Siliqi, 1993) was applied to calculated error-free data. It was shown that *ab initio*[*] crystal structure solution of proteins by direct methods is theoretically feasible if data from one isomorphous derivative are available.

In paper II, a new phasing method was proposed for the *ab initio* crystal structure solution of proteins. The method is based on a probabilistic approach which integrates direct methods and isomorphous techniques. The keystone is the formula estimating three-phase invariants given six magnitudes, obtained by Giacovazzo, Cascarano & Zheng (1988) (see also a related formula by Hauptman, 1982). Important points of the method are: (1) a correct choice of the reflections actively used in the phasing process (incorrect selection may not satisfy the statistical solvability criterion); (2) a normalization procedure that reduces the influence of the measurement errors and of the lack of isomorphism; (3) a correct weighting scheme in the tangent-refinement process designed for driving phases far away from the so-called 'Patterson solution'; (4) efficient new figures of merit for picking out the correct solution among several trials.

The experimental data of four proteins, quoted in Table 1 by code names APP, CARP, E2, M-FABP, were used. In Table 2, the key parameters of the phasing process are shown: DERIVATIVE denotes the heavy atom added to the protein, RES = $\lambda/(2\sin\theta_{\rm max})$ is the resolution of the measured data for the derivative, NREFL is the number of measured symmetry-independent reflections up to RES resolution, NLAR is the number of structure factors phased by the procedure, ERR is the weighted average phase error (°) for the NLAR assigned phases. The results may be described thus:

(*a*) NLAR reflections were phased in a straightforward way by a default run of the program. Only 25 trials were necessary for obtaining the correct solution. This number is astonishingly small if one considers the complexity of the problem.

(*b*) Figures of merit efficiently ranked the trial solutions: the correct solution was always among the first three.

(*c*) The procedure was not time consuming: 4 or 5 min of CPU time on an IBM risk 6000 were sufficient for the completion of the phasing process.

(*d*) Atomic resolution was not necessary. Data up to the derivative resolution were used.

(*e*) The procedure benefits by perfect isomorphism and accurate measurements but it is not very sensitive to lack of isomorphism and/or to experimental errors. CARP and M-FABP can each be considered as a good representative of a standard protein (with respect to quality of data, size *etc.*) while E2 has an excellent isomorphous derivative.

(*f*) Correlation between the electron-density maps calculated by using the NLAR reflections phased by our procedure and the maps relative to all the NREFL reflections with their true phases was high. However, our maps were not immediately interpretable because of: (1) too large series-truncation effects (*i.e.* NLAR was too small); (2) possible loss of enantiomorph (for APP and CARP). The general conclusion was that phase extension rather than better refinement of the assigned phases was the most urgent problem to solve.

[*] As usual for direct methods, we speak of *ab initio* crystal structure solution when phases are directly derived from diffraction data without any supplementary prior information on heavy-atom positions, orientation of the molecule, single isomorphous replacement or multiple isomorphous replacement *etc.*

Table 1. *Code name, space group and crystallo-chemical data for test structures*

Structure code	Space group	Molecular formula	Z
APP*	$C2$	$C_{190}N_{53}O_{58}Zn$	4
CARP†	$C2$	$C_{513}N_{131}O_{121}Ca_2S$	4
E2‡	$F432$	$C_{1170}N_{310}O_{366}S_7$	96
M-FABP§	$P2_12_12_1$	$C_{667}N_{170}O_{261}S_3$	4

* Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983).

† Kretsinger & Nockolds (1973).

‡ Mattevi, Obmolova, Schulze, Kalk, Westphal, De Kok & Hol (1992).

§ Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992).

Table 2. *Key parameters of the phasing process*

DERIVATIVE denotes the heavy atom added to the protein, RES $[=\lambda/(2\sin\theta_{\max})]$ is the resolution of the measured data for derivative, NREFL is the number of measured symmetry-independent reflections, NLAR is the number of largest normalized structure factors phased by the procedure described in paper II, ERR is the weighted average phase error for the NLAR assigned phases.

	DERIVATIVE	RES	NREFL	NLAR	ERR (°)
APP	Hg	2.0	2086	600	41
CARP	Hg	2.0	4416	1000	41
E2	Hg	3.0	7757	1000	30
M-FABP	Hg	3.0	2931	800	45

In accordance with the conclusions of paper II, the problem of how to preserve the enantiomorph in the phasing process will be treated in the next paper; we devote the present paper to extending phases to a larger number of reflections. It will be shown that phase extension can be performed without reducing the quality of the new phase values provided that a limited percentage of the total number of reflections measured up to RES resolution are involved in the procedure. The entire process, phase assignment and phase extension, can be fully automated.

Usefulness of indicators predicting errors in triplet estimation

Although all reflections may be involved in the phase-extension process, only a subset of structure factors can be phased reliably. The first problem to solve is: how many reflections can we involve in the phase-extension process without a strong reduction of the phase reliability?

Crick & Magdoff (1956) first established the usefulness of a parameter that measures the average change in intensity due to the addition of heavy atoms to a protein. More recently, a strictly connected parameter, the so-called diffraction ratio,

$$\mathrm{DR} = \{2[\sigma_2]_H/[\sigma_2]_p\}^{1/2} = (2\sigma)^{1/2},$$

has been employed by Fortier, Weeks & Hauptman (1984) for predicting the overall reliability of the phase estimates *via* direct methods. For exceedingly large values of the diffraction ratio, the integration of direct methods with isomorphous-replacement techniques produces marginal benefit; on the other hand, too small a value does not provide a sufficient signal-to-noise ratio. Detailed prior knowledge of DR is not needed for the estimation of triplet invariants: small errors in its estimate are not prejudical for direct-methods applications. A statistical evaluation of DR may be obtained *via* the correlation coefficient (Hauptman, 1982; see also Kyriakidis, Peschar & Schenk, 1993, for a related expression)

$$r = \langle (R^2 - \langle R^2\rangle)(S^2 - \langle S^2\rangle)\rangle \times \langle (R^2 - \langle R^2\rangle)^2\rangle^{-1/2}\langle (S^2 - \langle S^2\rangle)^2\rangle^{-1/2}, \quad (1)$$

where the averages are taken over all reciprocal-lattice vectors and $r \simeq 1/(1 + \mathrm{DR}^2/2)$.

If the averages in (1) are taken over all reciprocal-lattice vectors having a fixed value of $\sin\theta/\lambda$ then r is expected to be constant in the case of perfect isomorphism, monotonically decreasing with $\sin\theta/\lambda$ in the case of imperfect isomorphism. Unfortunately, the uncertainty in the relative scaling of the native and derivative intensity data does not allow an accurate estimate both of DR and of the degree of isomorphism. In Fig. 1, we plot r as a function of $\sin\theta/\lambda$ for APP, CARP, E2 and M-FABP. While r does not significantly vary for APP and M-FABP (this last is not characterized by a perfect isomorphism), it decreases at high $\sin\theta/\lambda$ for E2 and CARP, which are characterized by a quite good derivative and by a bad derivative, respectively.

Even if DR and the quality of the isomorphism were known *a priori*, they would not answer the problem of evaluating how many reflections should be involved in the phase-extension process. A useful suggestion may be derived as follows. According to paper I [equation (11)], the reliability parameter for a triplet phase is

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

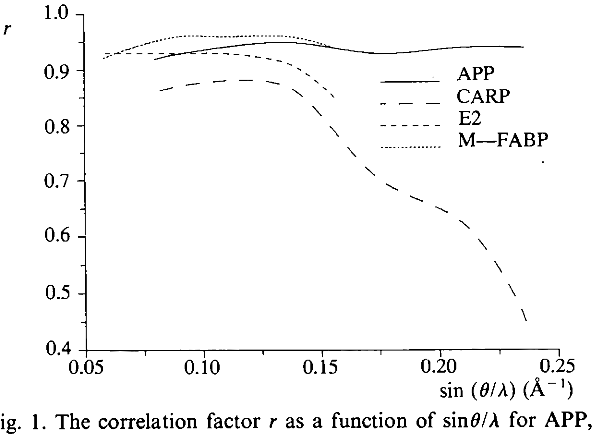


Fig. 1. The correlation factor r as a function of $\sin\theta/\lambda$ for APP, CARP, E2 and M-FABP.

For typical derivatives and for R and S larger than or close to unity, the factor T is so close to unity that Δ' may replaced by Δ. Relation (2) suggests that the overall reliability of the triplet phase estimates, and therefore the efficiency of the phasing process, is correlated with the distribution of $|\Delta|$. Thus, it is of interest to derive it and to exploit such information to guess about the phase-extension process.

The probability distribution function $P(|\Delta|)$

According to Hauptman (1982),

$$P(R,S) = [(4RS)/(1-\alpha^2)]\exp - [(R^2+S^2)/(1-\alpha^2)] \times I_0[(2\alpha RS)/(1-\alpha^2)], \quad (3)$$

where

$$\alpha \simeq \{[\sigma_2]_p/[\sigma_2]_d\}^{1/2}.$$

We first express (3) in terms of the pseudo-normalized structure factors S' and R',

$$P(R',S') = 4R'S'(\Sigma_H/\Sigma_p) \times \exp\{-[R'^2(\Sigma_d/\Sigma_p) + S'^2]\} \times I_0(2R'S'). \quad (4)$$

Then we introduce the change of variable $\Delta = S' - R'$ and (4) becomes

$$P(R',\Delta) = 4(\Sigma_H/\Sigma_p)R'(R'+\Delta) \times \exp\{-[2R'^2 + R'^2(\Sigma_H/\Sigma_p) + 2R'\Delta + \Delta^2]\} \times I_0[2R'(R'+\Delta)]. \quad (5)$$

For $-3.75 \leq x \leq 3.75$, $I_0(x)$ may be approximated by a polynomial in even powers of t (see Abramowitz & Stegun, 1972), where $t = x/3.75$. For large values of R' it is not easy to compute (5) directly. For $3.75 < x < \infty$, we approximate $I_0(x)$ by $Q(t)\exp(x)x^{-1/2}$, where Q is a suitable polynomial of order 8 in terms of t^{-1}.

We obtain

$$P(R',\Delta) = 2(2)^{1/2}(\Sigma_H/\Sigma_p) \times \exp(-\Delta^2)[R'(R'+\Delta)]^{1/2} \times Q(t)\exp[-R'^2(\Sigma_H/\Sigma_p)]. \quad (6)$$

Then,

$$P(\Delta) = \int_0^\infty P(R',\Delta)\mathrm{d}R', \quad (7a)$$

for positive values of Δ, and

$$P(\Delta) = \int_{-\Delta}^\infty P(R',\Delta)\mathrm{d}R', \quad (7b)$$

for negative values of Δ (the limits of integration are because $R' = S' - \Delta$ has to be positive). Finally,

$$P(|\Delta|) = P(+|\Delta|) + P(-|\Delta|). \quad (8)$$

The distribution $P(\Delta)$ has been calculated by numerical methods: for simplicity, we have replaced Σ_H/Σ_p by σ. Curves corresponding to various values of σ are shown in Fig. 2. As expected, $P(\Delta)$ is not an even function. The range (0.46, 0.04) includes most of the σ's found in the literature for protein crystallography. The value $\sigma = 0.46$ is unusual and corresponds to APP, $\sigma = 0.09$, 0.08, 0.06 are the corresponding values for CARP, E2 and M-FABP, respectively. Curves in Fig. 2 do not strongly vary with σ but each of them is significantly shifted with respect to the others.

In Fig. 3, we show the distribution $P(|\Delta|)$ calculated for $\sigma = 0.08$; it can be considered a satisfactory approximation of the distribution of $|\Delta|$ for a typical protein. In Fig. 4, we show the cumulative distribution of (8), together with cumulative curves of the Wilson distribution,

$$P_1(R) = 2R\exp(-R^2),$$

for non-centrosymmetric space groups; and

$$P_{\bar{1}}(R) = (2/\pi)^{1/2}\exp(-R^2/2),$$

for centrosymmetric space groups. The figure shows that the percentage of reflections with $|\Delta|$ larger than

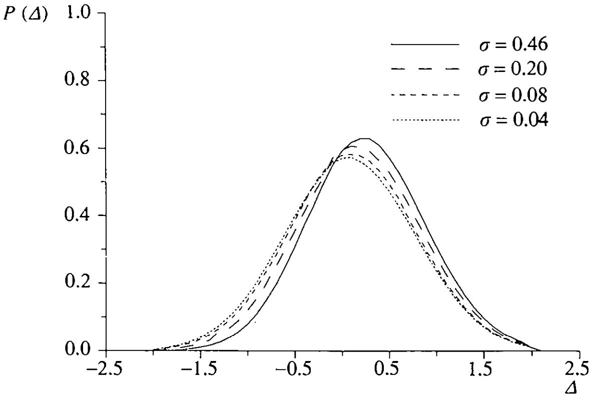


Fig. 2. $P(\Delta)$ distribution for selected values of σ.

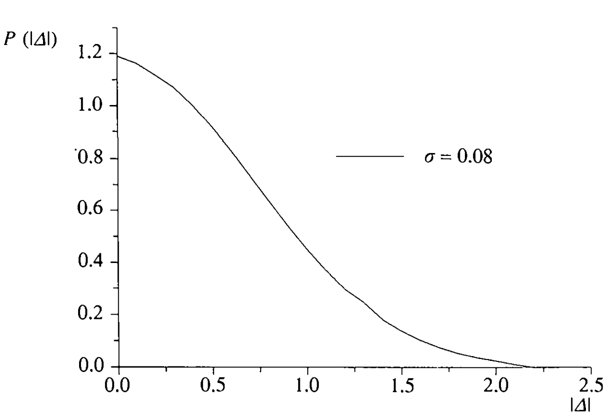


Fig. 3. $P(|\Delta|)$ for $\sigma = 0.08$.

a given threshold TRΔ (for example, TR$\Delta \simeq 0.5$) is significantly smaller than the corresponding percentage for the normalized structure factors in centro- and in non-centrosymmetric space groups. For the reader's convenience, numerical values for the cumulative distribution functions $P(|\Delta|)$ are given in Table 3.

What do these results suggest for the phasing process of the proteins? According to relationship (2), the triplet reliability mostly depends on the $|\Delta|$ parameters while the reliability of the single phase $\varphi_{\mathbf{h}}$ relies on the $\alpha_{\mathbf{h}}$ value [see equation (II.3)]. A basic condition for high $\alpha_{\mathbf{h}}$ values is that $|\Delta_{\mathbf{h}}|$ is sufficiently large: since $[\sigma_2^3/\sigma_3^2]_H$ is a very small number, $|\Delta_{\mathbf{h}}| > 0.5$ may be chosen (as a rule of thumb) as a reasonable lower limit for $|\Delta|$. If the rule is satisfied, $\varphi_{\mathbf{h}}$ is said to be *accessible* through the triplet relationships (I.11). According to Table 3, the number of accessible phases is about the 52% of NREFL, that is 1085, 2296, 4034 and 1524, for APP, CARP, E2 and M-FABP, respectively.

In paper II, the number of phased reflections (see the parameter NLAR in Table 2) was 600, 1000, 1000 and 800 for APP, CARP, E2 and M-FABP, respectively. They correspond to the ratios 'number of phased reflections/number of measured reflections' equal to 0.29, 0.23, 0.13 and 0.28, respectively. There is, therefore, the possibility of extending the phasing process to a substantial supplementary set of reflections. It is worthwhile mentioning (see papers I and II) that the experimental $|\Delta|$ values include a non-negligible noise as a consequence of errors in measurements, lack of isomorphism, scaling errors *etc.* The result is that the distributions $P(|\Delta|)$ and $C(|\Delta|)$ obtained from experimental data should be less sharp than expected theoretically. However, the results of this section suggest that, in ideal conditions, about 52% of the reflections up to the derivative resolution (*i.e.* those with the largest $|\Delta|$ values) could be phased by our probabilistic approach. The minimum value of $|\Delta|$ for these reflections is the threshold TRΔF which will be used in the phasing procedure. In the practical (non-ideal) cases, TRΔF can sometimes be smaller than 0.5, at which time we should be prepared to pay a penalty in terms of phase accuracy.

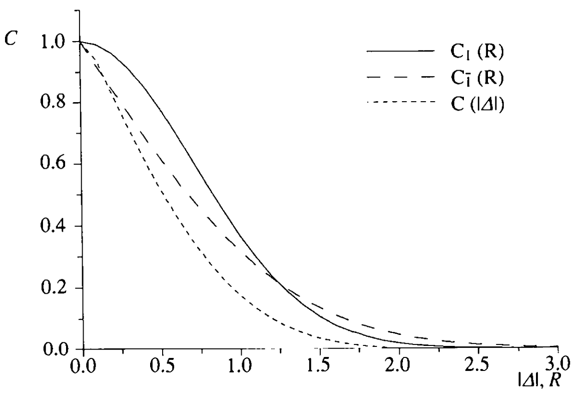


Fig. 4. Cumulative distribution of $P(|\Delta|)$ together with cumulative functions of Wilson distributions for centrosymmetric and non-centrosymmetric space groups.

Table 3. *Numerical values for the cumulative function* $C(|\Delta|)$

$\lvert\Delta\rvert$	%	$\lvert\Delta\rvert$	%
0.00	1.00	1.10	0.13
0.10	0.94	1.20	0.098
0.20	0.83	1.30	0.072
0.30	0.72	1.40	0.051
0.40	0.61	1.50	0.035
0.50	0.52	1.60	0.024
0.60	0.43	1.70	0.015
0.70	0.35	1.80	0.086
0.80	0.28	1.90	0.039
0.90	0.22	2.00	0.077
1.00	0.17		

The phasing procedure

In one possible strategy for phase determination, the threshold TRΔF for $|\Delta|$ could be fixed and all the reflections with $|\Delta| \geq$ TRΔF simultaneously involved in the phasing process. Triplets are then estimated and the tangent formula is applied. Such a strategy would require the calculation of several tens of millions of triplets, their cumbersome management by the tangent formula, and large storage and computing time.

We have chosen a different strategy: first we phase a small set of reflections with large $|\Delta|$ and R values. Among the various trials provided by a multisolution approach, the most probable ones will be used as seeds for subsequent phase expansion.

For the sake of simplicity, the procedure is described below in steps.

Step 1. *Selection of the reflections to phase*

As stated in papers I and II, the reflections to phase should be characterized by: (*a*) high values of $|\Delta|$, in order to guarantee a reliable phase assignment; (*b*) non-vanishing values of R, in order to provide, once phased, useful information for electron-density maps. Accordingly, the NREFL reflections (those for which both $|F_p|$ and $|F_d|$ are available from measurements) are partitioned into two subsets:

(1) The subset including the reflections with the smallest R values. Their number is chosen to be the minimum between 1000 and 25% of NREFL. Some of these reflections, *i.e.* those with $|\Delta| \leq 0.2$, will be used for constructing PSI0 triplets. Let NPSI be the number of reflections with small values of $|R|$ and $|\Delta|$ that are actually involved in PSI0 triplets.

(2) The subset $\{\gamma_1\}$ including NREFL-NPSI reflections. According to the preceding section, we should try to phase about 52% of the NREFL reflections, *i.e.* those characterized by the largest $|\Delta|$ values (accessible phases).

Step 2. *The first batch*

By default, 60% of the reflections in $\{\gamma_1\}$ (those with the largest R values) are selected. The cumulative distribution of the $|\Delta|$'s relative to such reflections is calculated, giving the number n of reflections with $|\Delta|$ larger than a given value. The threshold TRΔ(1) is chosen as the value of $|\Delta|$ corresponding to $n \simeq 800$. The *statistical solvability criterion* is applied: if it is satisfied then NLAR(1) = n is the number of reflections which will be phased first, otherwise NLAR(1) is increased until the criterion is satisfied. The NLAR(1) reflections are said to constitute the subset BATCH(1).

Step 3. *The next batch*

Let NLAR(2) be the number of reflections [among the NREFL-NPSI-NLAR(1) reflections] with $|\Delta| >$ TRΔ(2) $\equiv$ TRΔ(1). They will constitute the subset BATCH(2).

The remaining NREFL-NPSI-NLAR(1)-NLAR(2) reflections are divided into subsets [*i.e.* each BATCH(i) for $i > 2$ contains about 400 reflections], the ith subset being associated with a given threshold TRΔ(i) for $|\Delta|$. Since TR$\Delta(i+1) \leq$ TR$\Delta(i)$, the reflections in BATCH(i) will have $|\Delta|$ larger than the reflections in BATCH($i+1$). The last TRΔ value will coincide with TRΔF.

Step 4. *A supplementary batch*

In order to improve the continuity in the Fourier map, an additional number of reflections in the low $\sin\theta/\lambda$ range is phased. The corresponding subset [*i.e.* BATCH(LAST)] will involve reflections with $\sin\theta/\lambda \leq (\sin\theta/\lambda)_{\max}/2$, provided

$|\Delta| \geq \mathrm{TR}\Delta F \times 0.95 \times 0.85$ for reflections with restricted phase value,

$|\Delta| \geq \mathrm{TR}\Delta F \times 0.95$ for reflections of general type.

Step 5. *Triplet calculation*

Let $\{T_{ii}\}$ be the set of triplet invariants among the reflections in BATCH(i) and let $\{T_{ij}\}$ be the set of triplets constituted by one reflection in BATCH(i) and two reflections in BATCH(j), In our procedure, we only calculate the sets $\{T_{i1}\}$ for $i = 1,2,\ldots$ and we store for each ith set up to 50000 triplets (the most reliable ones).

Step 6. *The phasing procedure*

The NLAR(1) reflections in BATCH(1) are phased according to the procedure described in paper II. Among the various trials provided by the multisolution approach, the most probable one is chosen as a seed for the subsequent phase expansion.

The set BATCH(2) is phased from BATCH(1) by using the $\{T_{21}\}$ triplets: phases are then refined by making use of the triplets $\{T_{11}\} \cup \{T_{21}\}$. Since TR$\Delta$(2) $\equiv$ TRΔ(1), the average accuracy of the phases in BATCH(2) is expected to be very close to that of the reflections in BATCH(1). Therefore, for $i > 2$, the set BATCH(i) is phased from BATCH(1) by using the $\{T_{i1}\}$ triplets: phases are then refined by using the set of triplets $\{T_{11}\} \cup \{T_{21}\} \cup \{T_{i1}\}$. It is worthwhile noting that every set of phases so obtained is referred to the same origin, that fixed for set BATCH(1).

Step 7. *The Fourier map*

Once the set BATCH(1) ∪ BATCH(2) ∪ BATCH(3) ∪ ... is phased, it is used for calculating an electron-density map. If the map is not satisfactory (*i.e.* it is not interpretable in the chemical sense, then the trial immediately following [as ranked by the combined figure of merit (CFOM)] the most probable one is used as a seed for phase expansion (in accordance with step 6): the corresponding Fourier map is then calculated. The process may be cyclically repeated for each trial.

Applications

The procedure above has been applied to the four test structures quoted in Table 1. For each structure, only 20 trial solutions were demanded to our multisolution approach. The correct solution was that with the highest value of CFOM for M-FABP (CFOM = 0.43), with the second highest value for CARP (CFOM = 0.889), with the third highest value of CFOM for E2 and APP (CFOM = 0.425 and 0.954, respectively).

Below we will only give details about the phasing process for the correct solutions.

We first attempted extending phases up to 52% of the NREFL reflections. In order to check the quality of the assigned phases, we divided the reflections in ranges of $\sin\theta/\lambda$, α, R and $|\Delta|$: for each range, we calculated the phase error ERR just after the phase extension [ERR(1)] and after the tangent refinement [ERR(2)]. The results for M-FABP are shown in Table 4.

We observe:

(*a*) ERR(1) and ERR(2) are large at very small values of $\sin\theta/\lambda$. This is probably due to the scattering from the disordered water, which seems able to disturb the experimental estimates of $|\Delta|$ up to

Table 4. *Results for M-FABP*

The phasing process is extended to 1409 reflections, 1405 of which have non-zero weight after tangent refinement. The overall unweighted phase error is 52° after tangent refinement.

		NR(1)	ERR(1) (°)	NR(2)	ERR(2) (°)
(*a*)	$(\sin\theta/\lambda)^2$ (Å^{-2})				
	0.028–0.025	203	61	202	60
	0.025–0.019	446	54	445	53
	0.019–0.012	362	53	361	53
	0.012–0.006	256	45	255	44
	0.006–0.001	142	50	142	50
(*b*)	α				
	0–56	719	61	534	63
	56–112	261	50	386	53
	112–168	185	46	225	47
	168–225	92	39	106	36
	225–281	60	32	61	32
	281–337	34	25	35	24
	337–393	17	28	17	28
	393–449	15	28	15	28
	449–505	5	43	5	43
	505–562	12	42	12	42
	562–618	4	23	4	23
	618–674	2	19	2	19
	786–842	1	1	1	1
	898–955	1	0	1	0
	1067–1123	1	165	1	165
(*c*)	R				
	0.0–0.4	172	68	171	69
	0.4–0.8	496	57	493	56
	0.8–1.2	382	46	382	47
	1.2–1.6	221	50	221	50
	1.6–2.0	96	38	96	38
	2.0–2.4	30	43	30	43
	2.4–2.8	7	39	7	39
	2.8–3.2	4	59	4	59
	3.2–3.6	1	0	1	0
(*d*)	$\lvert\Delta\rvert$				
	0.0–0.4	11	82	11	82
	0.4–0.8	782	61	780	60
	0.8–1.2	448	44	447	45
	1.2–1.6	125	33	125	32
	1.6–2.0	26	22	25	16
	2.0–2.4	11	17	11	17
	2.4–2.8	3	0	3	0
	2.8–3.2	2	90	2	90
	3.2–3.6	1	165	1	165

Table 5. *APP*: *basic parameters of the phase-extension process*

BATCH	TRΔ	NLAR	ERR(2) (°)	
1	0.30	716	44	
2	0.28	90	64	
3	0.27	4	67	
	Total	810	46	ERR(weighted) = 43°

Table 6. *CARP*: *basic parameters of the phase-extension process*

BATCH	TRΔ	NLAR	ERR(2) (°)	
1	0.70	826	43	
2	0.70	400	47	
3	0.70	322	53	
4	0.64	169	57	
5	0.53	353	64	
6	0.44	41	52	
	Total	2111	50	ERR(weighted) = 46°

Table 7. *E*2: *basic parameters of the phase-extension process*

BATCH	TRΔ	NLAR	ERR(2) (°)	
1	0.90	912	27	
2	0.90	400	35	
3	0.90	250	40	
4	0.84	207	44	
5	0.77	238	38	
6	0.71	267	47	
7	0.65	281	52	
8	0.53	396	53	
9	0.53	181	57	
10	0.40	86	42	
	Total	3218	40	ERR(weighted) = 37°

Table 8. *M-FABP*: *basic parameters of the phase-extension process*

BATCH	TRΔ	NLAR	ERR(2) (°)	
1	0.45	709	46	
2	0.44	491	55	
3	0.36	31	60	
	Total	1231	50	ERR(weighted) = 47°

about 7–8 Å resolution. ERR increases for higher $\sin\theta/\lambda$: this is probably due to the progressive lack of isomorphism. A curious observation is that, owing to the imperfect isomorphism, direct methods can only work in practice at non-atomic resolution even if traditionally they are expected to work only at atomic resolution.

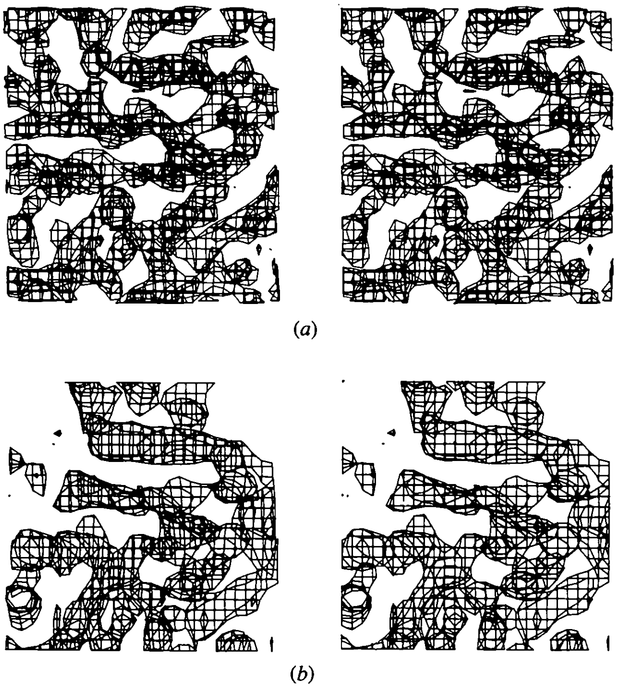
(*b*) ERR(1) and ERR(2) are strongly correlated with α. The effect of the phase refinement is substantially a rearrangement of the phase error as a function of α rather than its overall reduction. Thus, the present phase refinement is not indispensable.

(*c*) ERR(1) and ERR(2) are large for small values of R and small for large values of R. This behaviour can be explained thus: (i) small values of R cannot give rise to a valuable Cochran contribution in relation (2); (ii) the standard deviation of the measured intensities is usually larger for weak reflections, and this can cause the accuracy of the experimental Δ values to deteriorate. Table 4(*c*) supports our procedure, according to which the NLAR(1) reflections (the first seed) are selected among the reflections in $\{\gamma_1\}$ with largest R values. Indeed, their phases must be as accurate as possible since the error quite easily propagates to the other batches.

(*d*) ERR is small for large values of $|\Delta|$ and large for small values of $|\Delta|$. This behaviour is just the expected effect of (2). It is worthwhile noting that our rule of thumb on the minimum value of $|\Delta|$ (*i.e.* $|\Delta_{\mathbf{h}}| > 0.5$) for a direct procedure is confirmed.

Trends in Table 4 were confirmed by analogous results for APP, CARP and E2. We then decided to modify slightly our procedure and declined to extend phases to reflections with $R < 0.4$. The penalty should not be too great since these reflections shold not contribute much to the electron-density map.

The modified procedure will now phase less than 52% of NREFL but their phase accuracy will probably improve. The number of phased reflections is 810, 2111, 3218 and 1231 for APP, CARP, E2 and MFABP, respectively, corresponding to 0.39, 0.48, 0.41 and 0.42 of NREFL. The reader can follow the phase-expansion and refinement processes through Tables 5–8, where for each BATCH the threshold value TRΔ, the number of phased reflections NLAR and the phase ERR(2) are shown. As expected, the phase error increases for the batches of high order but the overall error at the end of the procedure is still acceptable. The unweighted error passes for APP from 44° for the first seed to 46° for the 810 phased reflections, from 43 to 50° for CARP, from 27 to 40° for E2 and from 46 to 50° for M-FABP.

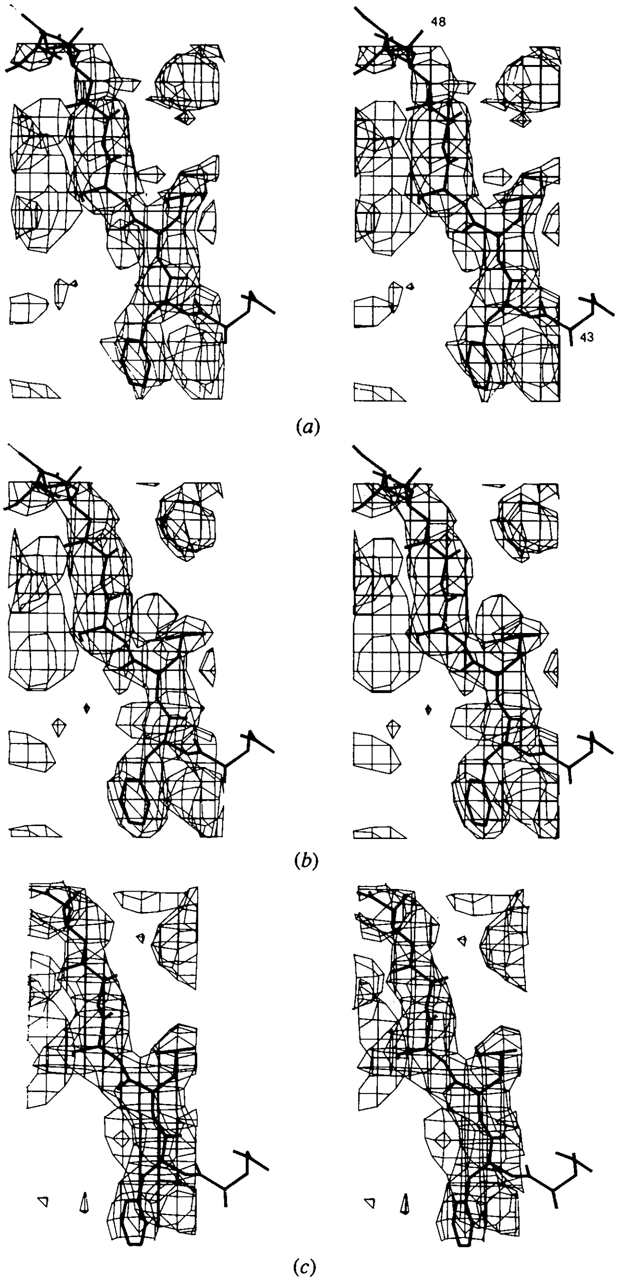

(*a*)

(*b*)

Fig. 5. Stereo drawing of the electron-density map of a portion of the crystal cell of E2, calculated at 6 Å resolution with (*a*) direct-methods phases and (*b*) phases derived from SIR and solvent flattening. The map is viewed along *z*, from 35 to 49 Å. *x* runs horizontally across the page, from 9 to 58 Å and *y* vertically, from 0 to 52 Å. The origin is at the top left corner. C^{α} atoms of the asymmetric unit are included in these sections; *i.e.* residues 70–93, 160–169 and 174–185 are shown as thin lines. [All the maps are calculated with the program *X-PLOR*3.0 (Brünger, Kuriyan & Karplus, 1987) and displayed on an Evans and Sutherland PS300 with the program *FRODO* (Jones, 1978). The SIR phase calculation and the solvent-flattening procedure we carried out using the program *PHASES*.]

The electron-density maps

In order to reduce the noise in the electron-density maps, a correct weight should be associated with each reflection. Table 4(*b*) shows that the α range can be extremely wide: therefore, a weight directly



(*a*)

(*b*)

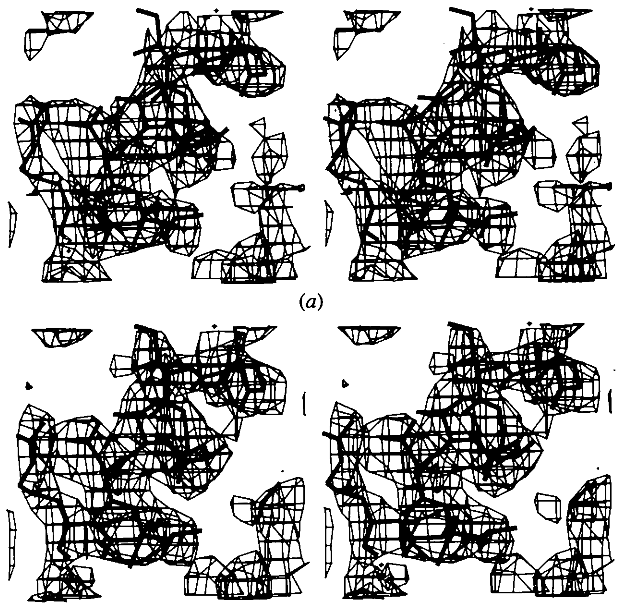
(*c*)

Fig. 6. Stereo drawings of a portion of the electron-density map of E2 around residues 43–49. Maps were calculated at 3 Å resolution with (*a*) direct methods, (*b*) model phases, using the same number of reflections (3226) and (*c*) phases derived from SIR and solvent flattening, using all the reflections that could be phased using the derivative (7751).

proportional to α would lead to small weighted phase errors [ERR(weighted) = 39, 42, 31 and 43° for APP, CARP, E2 and M-FABP, respectively] but could practically make negligible the contribution to the electron-density map of a large percentage of reflections (this weighting scheme was used in paper II). On the other hand, the weight $w_{\mathbf{h}} = D_1(\alpha_{\mathbf{h}})$, largely used for small-molecule electron-density maps, is also unsuitable: indeed, too high a percentage of reflections may have $w = 1$ (see Table 4*b* again), and this closely corresponds to the unweighted situation. We have preferred to use for the electron-density-map calculation a weighting scheme according to which 10% of the reflections (those with largest α values) have unitary weight: the other reflections are weighted by $w = D_1[f]$, where f is a smooth function increasing with α and $(\sin\theta/\lambda)^{-1}$. Weighted errors are shown in Tables 5–8 for each structure.

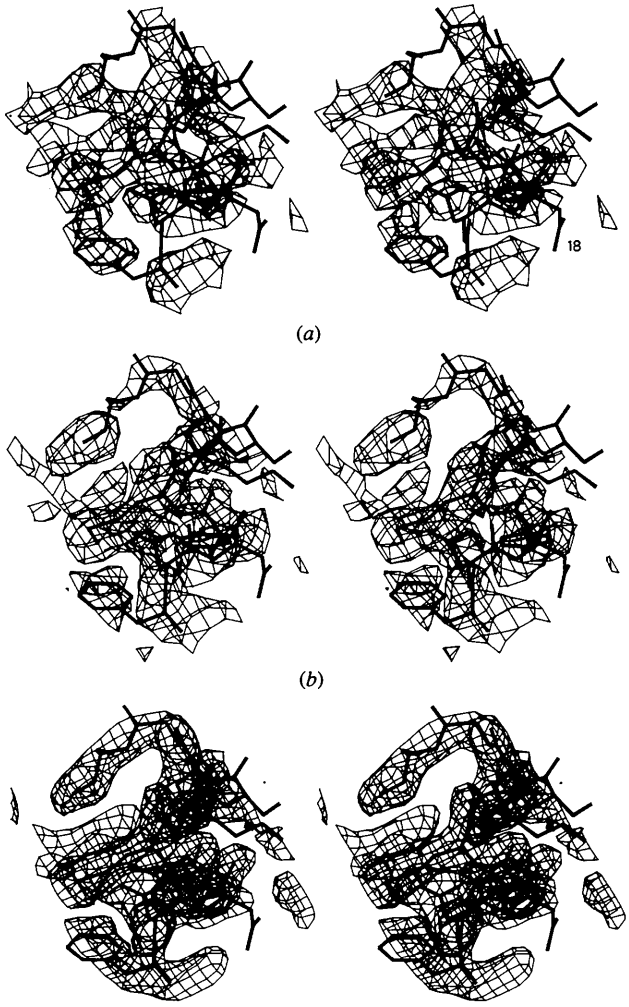
Since enantiomorphism is lost in APP and CARP, we calculated electron-density maps of E2 and M-FABP. Portions of the electron-density maps of E2 and M-FABP are shown in Figs. 5–8. A portion of the map calculated for E2 at 6 Å resolution with direct-methods phases is shown in Fig. 5, maps at 3 Å resolution are shown in Figs. 6 and 7. When useful, they are compared with corresponding maps obtained *via* model phases or with 'solvent-flattening' phases. The latter were obtained starting from SIR phases, after applying an automatic solvent-flattening procedure as implemented in the *PHASES* program (W. Furey, VA Medical Center and Univ. of Pittsburgh, PA, USA). Direct-methods maps are satisfactory: they are virtually identical to those obtained with model phases using the same number of reflections. Moreover, the electron density shows all the features of the molecular model.

Fig. 5(*a*) shows a large portion of a 6 Å-resolution map: the map is quite continuous, demonstrating that direct-methods phases are quite reliable even at very low resolution. The corresponding map obtained by the solvent-flattening procedure applied to SIR phases is shown in Fig. 5(*b*). The two maps are of comparable quality in spite of the fact that the latter is calculated with a number of reflections which is nearly double that used by direct methods.


(*a*)

(*b*)

Fig. 7. Same as Fig. 6, for density around residues 24–33.



(*a*)

(*b*)

(*c*)

Fig. 8. Stereo drawings of the electron-density maps of M-FABP around α-helix I, residues 16–25. All maps are calculated at 3 Å resolution, using (*a*) direct-methods phases, (*b*) model phases with the same number of reflections as the previous one and (*c*) model phases and *all* the reflections to 2.1 Å resolution.

The solvent-flattening process, however, improves the phases by a more careful recognition of the solvent regions in the electron-density map.

In Fig. 6, not only the envelope of the density is continuous, but also side chains, *e.g.* Phe44 and Glu45, are clearly distinguishable. Phases obtained by the solvent-flattening procedure (Fig. 6*c*) provide essentially the same information as direct-methods phases, even if they improve contrast between the protein and the solvent region. In Fig. 7, a portion of chain clearly presents the features of a α-helix. These results are quite surprising if we consider that we have used in the calculations only about a half of the reflections at the derivative resolution. Nevertheless, they can be explained if we consider that the reflections we are using are in general the strongest ones and the overall mean error on the phase angles is 40°.

A correlation coefficient between the published molecular model and the electron-density map was calculated according to Jones, Zou, Cowan & Kjeldgaard (1991) as implemented in the *O* program. The function, which takes values from -1 to 1, was calculated for the main-chain atoms of every residue. It usually gives information about the quality of the model, but in our case was used to extrapolate information about the quality of our map assuming the model is correct. In Fig. 9, the correlation coefficient for the molecular model of E2 is plotted against the number of residues for the direct-methods map (map *A*), for the map calculated with the same number of reflections and model phases (map *B*), and for the map calculated with 3 Å data and model phases (map *C*). The first two compare very well, *i.e.* the quality of our map (A) is practically the same as for map *B*. Small breaks in the density are present in both maps, and the overall correlation coefficient is 0.476 and 0.594 for maps *A* and *B*, respectively. The quality of the 3 Å map (map *C*) is obviously better, being based on phases obtained from the final refined model: no breaks are present in the main-chain density but the overall coefficient, 0.698, does not differ strongly from the previous two values. This situation is illustrated in Fig. 10, where a large portion of the three maps is reported.

The same considerations do not apply for M-FABP: the electron density calculated using 1400 reflections presents most of the correct features but is poor if compared with the map calculated with the same number of reflections and model phases (Fig. 8).

Several breaks are present in the main-chain envelope and a noise level is present that makes it difficult to build a molecular model. A comparison of this map with those reported in paper II of this series suggests that more or less the same details are revealed.

Concluding remarks

The set of phases provided by the phasing method described in paper II was not sufficiently large to produce interpretable electron-density maps. The procedure described here extends phases up to about

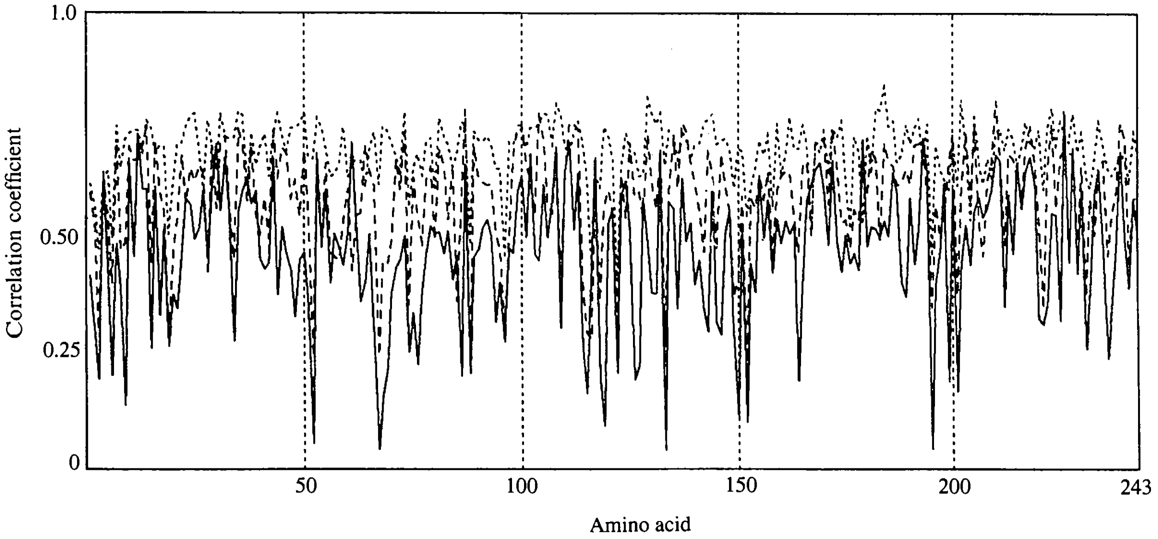


Fig. 9. Correlation coefficients between the model and the electron-density map, calculated with the program *O* (Jones, Zou, Cowan & Kjeldgaard, 1991). The continuous line represents the map calculated with phases from the present paper. The dashed line represents the map calculated with model phases and the same number of reflections (3226). The short-dashed line represents the map calculated with model phases using all the reflections at 3 Å resolution (8135).

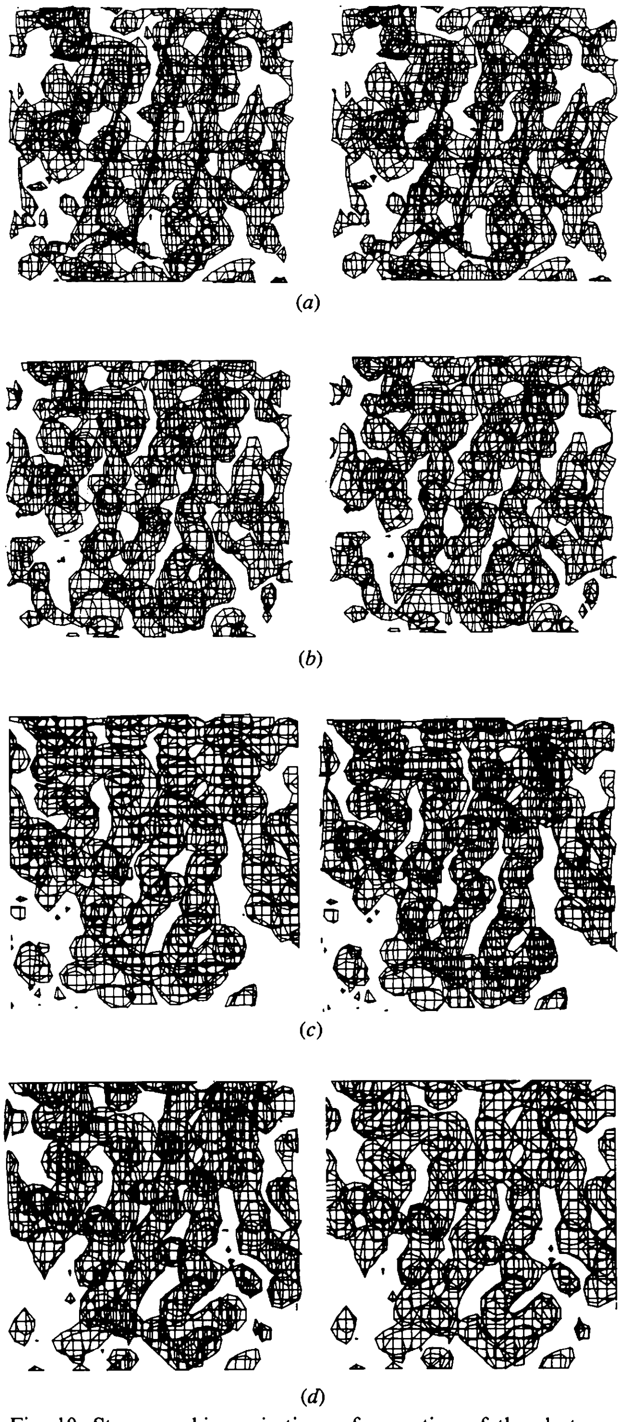


Fig. 10. Stereographic projections of a portion of the electron-density map E2 viewed along y (x extending from 18 to 48 Å, y from 18 to 28 Å and z from 24 to 54 Å) for the following maps: (a) phases from the present paper, 3226 reflections up to 3 Å resolution; (b) phases from the refined model, same number of reflections as in (a); (c) phases from the refined model, 8135 reflections up to 3 Å resolution; and (d) phases from SIR and solvent flattening, 7751 reflections. In map (a), C^{α} atoms for all the residues of one asymmetric unit that belong to these sections (amino acids from 43 to 46, 110 to 124, 158 to 173, 181 to 192, 208 to 214 and 221 to 228) are drawn.

40% of the reflections (up to the derivative resolution). The process is fast, does not require the calculation and the simultaneous use of millions of triplets and may be run in a completely automatic way. Thus, thousands of phases can be available with negligible computing time.

It is evident that the phase-extension procedure has produced an easily interpretable map in the case of E2 and has partially failed in substantially improving the map of M-FABP. This could be mainly ascribed to two effects: in the former case, phases have been extended from 1000 to 3200 reflections; in the latter, only 400 reflections have been added to the original 800. That is, the number of new reflections added for M-FABP has increased the set by about 50%, while the reflections added to E2 represent about twice the original number.

It is worth noting that the phasing process relies on a heavy-atom derivative and its quality has a strong influence on the final result: the heavy-atom derivative of M-FABP was not perfectly isomorphous and this increases the phase error. However, the mean phase error is increased more in the former case, from 27 to 40°, than in the latter, from 46 to 50°.

Finally, the following drawbacks still limit the usefulness of the present phase-extension process.

(1) Even if the number of phased reflections is sufficiently large for practical purposes, a non-negligible number of reflections with $|\Delta| \simeq 0$ but large R values remain unphased [phase inaccessible *via* relation (2)]. They could provide, once phased, a valuable contribution to the electron-density map. This extension process could also be obtained eventually by a 'solvent flattening'-like procedure or other technique complementary to that described in this paper, and could eventually help in all cases, like the M-FABP, where phases are substantially correct but insufficient to give the final solution.

(2) The overall phase error is moderately large. Its eventual reduction should allow a better definition of the protein envelope.

(3) The phase-refinement process following the phase-expansion procedure is fast but inefficient. This is probably because we only use the subsets $\{T_{i1}\}$ of the entire family of triplets.

(4) Pseudo-centrosymmetrical phases are provided in specific space groups.

Our next efforts will be devoted to overcoming the above points.

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover Publications, Inc.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 258–460.

CRICK, F. H. & MAGDOFF, B. (1956). *Acta Cryst.* **9**, 901–908.
FORTIER, S., WEEKS, C. M. & HAUPTMAN, H. (1984). *Acta Cryst.* **A40**, 544–548.
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**, 503–510.
GIACOVAZZO, C., SILIQI, D. & SPAGNA, A. (1994). *Acta Cryst.* **A50**, 609–621.
GLOVER, I., HANEEF, I., PITTS, J., WOODS, S., MOSS, D., TICKLE, I. & BLUNDELL, T. L. (1983). *Biopolyers*, **22**, 293–304.
HAUPTMAN, H. (1982). *Acta Cryst.* **A38**, 289–294.
JONES, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.
JONES, T. A., ZOU, J.-Z., COWAN, S. W. & KJELDGAARD, M. (1991). *Acta Cryst.* **A47**, 110–119.
KRETSINGER, R. H. & NOCKOLDS, C. E. (1973). *J. Biol. Chem.* **248**, 3313–3326.
KYRIAKIDIS, C. K., PESCHAR, R. & SCHENK, H. (1993). *Acta Cryst.* **A49**, 350–358.
MATTEVI, A., OBMOLOVA, G., SCHULZE, E., KALK, K. H., WESTPHAL, A. H., DE KOK, A. & HOL, W. G. J. (1992). *Science*, **255**, 1544–1550.
ZANOTTI, G., SCAPIN, G., SPADON, P., VEERKAMP, J. H. & SACCHETTINI, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.

Acta Cryst. (1995). A**51**, 188–197

The Estimation of Crystal Thickness and the Restoration of Structure-Factor Modulus from Electron Diffraction: a Kinematical Approach

BY D. TANG, J. JANSEN AND H. W. ZANDBERGEN

National Centre for HREM, Laboratory of Materials Science, Delft University of Technology, Rotterdamseweg 137, 2628 *AL Delft, The Netherlands*

AND H. SCHENK

Laboratory for Crystallography, University of Amsterdam, Nieuwe Achtergracht 166, 1018 *WV Amsterdam, The Netherlands*

(*Received* 6 *April* 1994; *accepted* 8 *September* 1994)

Abstract

A technique of crystal-thickness estimation and structure-factor-modulus restoration (reconstruction) from electron diffraction patterns, for use in crystal-structure determination, is proposed based on the kinematic scattering theory. A criterion for a self-consistent test of the restored structure-factor modulus has also been introduced from the structure-factor statistics developed by direct methods for X-ray diffraction. Theoretical tests on some structures are successful and show that the diffraction intensities are improved to be closer to the moduli of the true structure factors.

1. Introduction

The techniques of combined high-resolution electron microscopy (HREM) with electron diffraction intensity have been used for both HREM image deconvolution and resolution enhancement (*e.g.* Ishizuka, Miyazaki & Uyeda, 1982; Fan, Zhong, Zheng & Li, 1985; Liu *et al.*, 1990; Downing, Meisheng, Wenk & O'Keefe, 1990; Dong *et al.*, 1992; Hu, Fan & Li, 1992; Zou, Hovmöller, Parras, González-Calbet, Vallet-Regí & Grenier, 1993). These techniques are very useful in cases when crystals are too small for X-ray or neutron diffraction. Nearly all of these studies were for the kinematical condition (weak-phase-object approximation) or near the kinematical condition (pseudo-weak-phase-object approximation) (Tang & Li, 1988), that is when electron dynamical scattering is not predominant. Under such conditions, the phase of the diffracted wave function is replaced by the phase of the Fourier transform of the corresponding high-resolution electron-microscope image so that the phase problem that occurs in X-ray diffraction can be partly resolved.

Although the dynamical-diffraction effect is much stronger in electron diffraction than in X-ray diffraction, the dynamical perturbations to the diffracted beams are expressed as phase distortions before the wave amplitudes change much from their kinematical values (Dorset, Tivol & Turner, 1992). That is to say, the electron diffraction intensity is proportional to the square of the modulus of the structure factor in a greater range of thickness than that for which kinematical diffraction is valid.

A well known formula for the kinematical diffracted intensity, neglecting the Lorentz–polarization correction, gives the relative intensity as (Vainshtein, 1964; Cowley, 1988)

$$I(\mathbf{g}) = |F(\mathbf{g})|^2[(\sin \pi s_g t)/\pi s_g t]^2. \qquad (1)$$