*Acta Cryst.* (1994). A**50**, 609–621

# The *Ab Initio* Crystal Structure Solution of Proteins by Direct Methods. II. The Procedure and its First Applications

By Carmelo Giacovazzo and Dritan Siliqi*

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

and Riccardo Spagna

*Istituto di Strutturistica Chimica 'G. Giacomello', CNR, Area della Ricerca, CP* 10,
00016 *Monterotondo Stazione, Roma, Italy*

(*Received* 24 *November* 1993; *accepted* 28 *February* 1994)

## Abstract

A direct phasing method is described that is potentially able to solve *ab initio* protein structures. The method uses the information contained in diffraction data of the native structure and of one isomorphous derivative. The various steps of the procedure are analysed in order to estimate their robustness against experimental errors in measurements and lack of isomorphism. Experimental tests involve four typical protein structures and show that crystal structure solution is attained in a rather straightforward way.

## Symbols and abbreviations

See paper I (Giacovazzo, Siliqi & Ralph, 1994).

## Introduction

In paper I of this series, it was concluded that *ab initio* crystal structure solution of proteins by direct methods is theoretically feasible if data from one isomorphous derivative are available. The statement was the consequence of the application of the *statistical solvability criterion* (Giacovazzo, Guagliardi, Ravelli & Siliqi, 1994) to calculated error-free data. It was shown that success can be attained if reflections used in the phasing process are characterized by large values of $R$ and $|\Delta'|$. Even if that conclusion was supported by statistical calculations on triplet invariant reliability, the feasibility of *ab initio* direct phasing was not proved in practice. Indeed, lack of isomorphism between native and derivative structures combined with errors in the experimental data and/or in their mathematical treatment could hinder success in practice (*i.e.* when the phasing procedure is applied to experimental data) even if the structure solution is straightforward with ideal error-free data. Generally, a direct procedure is unsuitable for practical purposes if it is too sensitive to data resolution, structural complexity, errors in the data and lack of isomorphism. A robust procedure should provide a reasonable rate of success in standard situations, for example when applied to data at nonatomic resolutions and with an accuracy level attainable by current experimental techniques. In particular, a phasing procedure successfully working with 3 Å resolution data would certainly solve many of the problems in macromolecular crystallography. We propose in this paper a phasing method that, applied to real experimental data, shows important features: it works at atomic and nonatomic resolutions, it is not sensitive to structural complexity and it is able to handle standard-quality data and imperfect isomorphism. The limits and the first applications of the method are also described. For clarity, the phasing procedure is described step by step.

We use as test data the experimental data of the four proteins quoted in paper I. APP data were collected using a four-circle diffractometer for the native and an $HgCl_2$ derivative (Blundell, Pitts, Tickle, Wood & Wu, 1981). The structure was solved by applying SIRAS (single isomorphous replacement including anomalous scattering) techniques to 2 Å resolution data. Phases were extended to 1.4 Å resolution by using a modified tangent formula. New data for the native protein up to 0.98 Å resolution were collected by a four-circle diffractometer (Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell, 1983).

For CARP (carp muscle calcium-binding protein), isomorphous and anomalous scattering data were measured (Kretsinger & Nockolds, 1973) up to 2.0 Å resolution using precession photography; three

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

heavy-atom derivatives were used. In our calculations, we only make use of the (3-chloromercurio-2-methoxypropyl)urea (CMMPU) derivative.

Diffraction data of E2 (catalytic domain of *Azotobacter vinelandii* dihydrolipoyl transacetylase) were collected on a fast television area detector (Mattevi, Obmolova, Schulze, Kalk, Westphal, De Kok & Hol, 1992). One mercury and two platinum derivatives were used for phasing: data included anomalous-dispersion effects [multiple isomorphous replacement including anomalous scattering (MIRAS)]. We only make use of the mercury derivative, which, as stated by Mattevi *et al.*, is of excellent quality.

The structure of M-FABP (recombinant human-muscle fatty-acid-binding protein) was originally solved using both multiple isomorphous replacement and molecular replacement procedures (Zanotti, Scapin, Spadon, Veerkamp & Sacchettini, 1992). Data for native and two isomorphous derivatives were collected with a Siemens X1000 area detector system and on a SDMS area detector system coupled with a rotating-anode generator. For our calculations, we used the $HgAc_2$ derivative.

Anomalous dispersion is neglected in our calculations: we use the approximation $F=(F^+ + F^-)/2$ (this introduces a supplementary error in the data).

Therefore, one small (APP) and three usual-size proteins are among the test structures. Furthermore, E2 has an unusually good derivative but low-resolution data and CARP is an example of particularly unfavourable data (imperfect isomorphism of the derivative, old experimental technique for data collection).

Even if the procedure is devised for and is applied to experimental data, some calculations are made with ideal error-free data in order to show the different response of the phasing process to different qualities of data. This will also enable the reader to better understand the effects of the various sources of errors on the various steps of the procedure and their overall influence on the robustness of the process.

## The normalization process

Several sources of errors affect the accuracy of the scale ($K$) and temperature ($B$) factors provided by the Wilson method (Hall & Subramanian, 1982; Cascarano, Giacovazzo & Guagliardi, 1992*a*). Compared with small-molecule crystallography, additional sources of errors in macromolecular crystallography are nonatomic resolution of data, disordered water distribution and possible lack of information concerning number and chemical occupancy of the crystallographic sites accommodating additional heavy atoms in derivative structures. The consequent errors for $K$ and $B$ could be critical for the success of direct procedures. Indeed, our probabilistic formula estimating triplet phase invariants [see equations (7) and (11) of paper I for the reliability parameter] mostly depends on $\Delta'$ factors which could be very sensitive to errors in $K$ and $B$.

Table 1. *Values of the scale and temperature factors obtained in the normalization procedure for error-free calculated data*

| Structure | | $(K_{sw})_p$ | $(B_{sw})_p$ | $(K_{Dw})_p$ | $(B_{Dw})_p$ |
|---|---|---|---|---|---|
| code (RES) | $B_{true}$ | $(K_{sw})_d$ | $(B_{sw})_d$ | $(K_{Dw})_d$ | $(B_{Dw})_d$ |
| APP | 6.8 | 1.27 | 6.0 | 0.87 | 9.2 |
| | | 0.91 | 9.1 | 0.91 | 9.1 |
| CARP | 18.3 | 0.95 | 18.40 | 0.96 | 18.3 |
| | | 1.00 | 18.1 | 1.00 | 18.1 |
| E2 | 20.0 | 0.66 | 28.1 | 0.67 | 27.1 |
| | | 0.69 | 27.1 | 0.69 | 27.1 |
| M-FABP | 20.0 | 0.85 | 20.6 | 0.76 | 24.3 |
| | | 0.81 | 24.4 | 0.81 | 24.4 |

In order to check the accuracy of the normalization process, we first applied the standard Wilson method to error-free calculated data for the test proteins used in paper I. For the structure-factor calculation, we chose the same isotropic temperature factor $B_{true}$ for all the atoms: the list of $B_{true}$ values chosen for the various test structures is shown in Table 1. Calculated data are obviously on an absolute scale ($K_{true} \equiv 1$). Standard Wilson plots applied to calculated protein data at the native protein resolution provided for $K$ and $B$ the values $(K_{sw})_p$ and $(B_{sw})_p$, respectively; when applied to calculated derivative data at the same resolution as the measured data of the derivative, the values $(K_{sw})_d$ and $(B_{sw})_d$ were obtained (see Table 1).

Deviations from true values are not negligible: as expected, any error in $B$ is correlated with a corresponding error in $K$ and *vice versa*. The worst situation occurs for APP, where the passage from 2 to 1 Å resolution involves a large difference between the reflection numbers used in the normalization process (2086 at 2 Å *versus* 17058 at 1 Å resolution).

It may be observed now that errors in the ratio $K_d/K_p$ and the difference $B_d - B_p$ are much more critical for the phasing process (they can invert $\Delta'$ signs) than errors in absolute values of $K$ and $B$. We therefore decided to first calculate $(K_{sw})_p$ and $(B_{sw})_p$ by the standard Wilson method at the derivative resolution and then estimate $K_d/K_p$ and $B_d - B_p$ by a differential Wilson plot through the equation (see Blundell & Johnson, 1976)

$$\ln[(\textstyle\sum_p + \sum_H)\langle F_p^2\rangle / \sum_p \langle F_d^2\rangle] = \ln(K_p/K_d) + 2(B_d - B_p)\sin^2\theta/\lambda^2.$$

The new values of $K$ and $B$ for derivatives are denoted $(K_{Dw})_d$ and $(B_{Dw})_d$ in Table 1. Again, errors are not negligible but the new estimates of $K_d/K_p$ and

$(B_d - B_p)$ are now more promising; indeed, they do not seem so large as to invert the signs of $\Delta'$ terms with the largest modulus. Accordingly, reflections with small $|\Delta'|$ values (as a rule of thumb, $|\Delta'| < 0.20$) should never be actively used in a phasing process if formulas (7) and (11) of paper I are applied. However, it is shown in the next sections that inversion of the $\Delta'$ terms can also occur for $\Delta' \gg 0.2$ as a consequence of the lack of isomorphism and of the errors in measurements combined with errors in scaling and thermal factors. Luckily, our statistical solvability criterion predicts that direct structure solution for macromolecules is feasible provided reflections with sufficiently large values of $R$ and $|\Delta'|$ are used in the phasing process. From now on, by 'normalization procedure' we will always mean a differential Wilson plot.

A supplementary tool for evaluating the general practical effectiveness of the normalization procedure may be the comparison between the $|\Delta'|$ distribution obtained by application of an ideal normalization procedure and error-free calculated data ($K = K_{\rm true}$, $B = B_{\rm true}$) with the $|\Delta'|$ distribution obtained by application of the differential Wilson plot to experimental data. Both distributions were calculated for the NLAR reflections (those with largest R value) defined in Table 4 of paper I. Ideal $|\Delta'|$ distributions for the test structures were shown in Fig. 3 of paper I and are here quoted in Fig. 1. The experimental $|\Delta'|$ distributions are shown in Fig. 2. Their comparison suggests two things.

(*a*) The APP curve in Fig. 2 is sharper than the corresponding curve in Fig. 1. The maximum $|\Delta'|$ value is about 1.4 for calculated error-free data and only 0.8 for observed data. Furthermore, the observed distribution shows too large a percentage of $|\Delta'|$ close to zero.

(*b*) The curves for CARP, E2 and M-FABP in Fig. 2 are flatter than the corresponding curves in Fig. 1 and show larger tails at the right-hand side. Consequently, a non-negligible percentage of $|\Delta|$'s are larger than their true values. If on one side too small $|\Delta'|$ values can perturb the phasing process, on the other side too large $|\Delta'|$'s might dominate it, making structure solution difficult. We therefore decided to rescale the experimental $|\Delta'|$ values in order to make their distribution closer to the expected one.

From the relation

$$|F_H|^2 = |F_d|^2 + |F_p|^2 - 2|F_pF_d| \cos q,$$

where $q$ is the angle between $F_p$ and $F_d$, the normalized expression

$$|E_H|^2 = |E'_d|^2 + |E'_p|^2 - 2|E'_pE'_d| \cos q \qquad (1)$$

is obtained, where $E_H$ is the normalized structure factor for the heavy-atom structure and $E'_d$ and $E'_p$ are pseudonormalized structure factors defined by

$$E'_d = F_d/\Sigma_H^{1/2}, \quad E'_p = F_p/\Sigma_H^{1/2}.$$

In accordance with Hauptman (1982), $\cos q$ may be approximated by $T_1 = D_1(2\beta_{oi}R_iS_i)$ (see paper I for an explanation of the symbols). Averaging over all reflections provides

$$\langle |E_H|^2 \rangle = 1 = \langle |E'_d|^2 + |E'_p|^2 - 2|E'_pE'_d|T_1 \rangle.$$

The scale factor

$$S = (\langle |E'_d|^2 + |E'_p|^2 - 2|E'_pE'_d|T_1 \rangle)^{-1/2}$$

is used for rescaling $|\Delta'|$ values.

The distributions of rescaled $|\Delta'|$'s for experimental data are shown in Fig. 3; they are significantly closer to the expected ones. For example: (*a*) the maximum value of $|\Delta'|$ for APP is 0.8 before rescaling and 1.5 after rescaling (to be compared with 1.4, the maximum value of $|\Delta'|$ for error-free data; (*b*) the maximum value of $|\Delta'|$ for E2 is about 1.8 for ideal

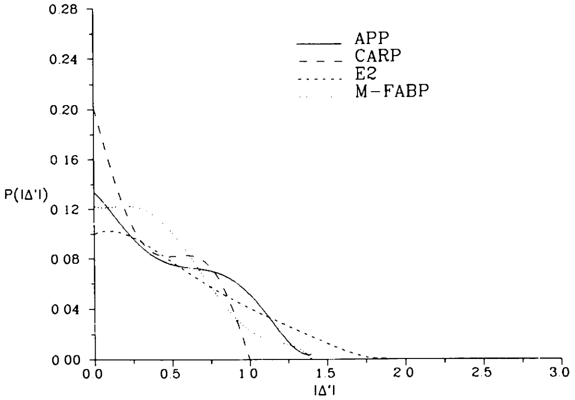

Fig. 1. $|\Delta'|$ distribution for error-free calculated data after an ideal normalization process ($K = K_{\rm true}$, $B = B_{\rm true}$). The distribution refers to the NLAR reflections.

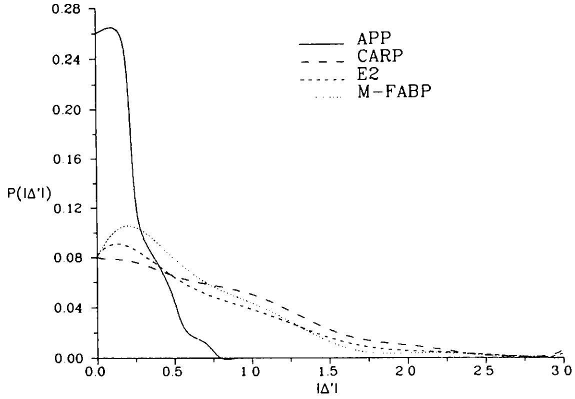

Fig. 2. $|\Delta'|$ distribution for experimental data after normalization by a differential Wilson plot. The distribution refers to the NLAR reflections.

error-free data; for observed data before rescaling there are four $|\Delta'|$ values larger than 3 and none for observed data after rescaling.

It is worthwhile noting that perfect correspondence between error-free and observed $|\Delta'|$ distributions will never be obtained. Indeed, lack of isomorphism and measurement errors can introduce remarkable noise in $\Delta$'s (see later parts of this paper). The worst situation occurs for CARP, where among the NLAR reflections there are 46 $|\Delta'|$'s larger than 2 before rescaling and 37 after rescaling. In our procedure, we introduce the supplementary criterion, according to which if $|\Delta'| > 2.0$ then $|\Delta'|$ is reset to 2.0. The $|\Delta'|$ values so modified act as observed in the next steps of the procedure.

## The phasing procedure

According to the suggestions of paper I, the reflections actively used in the phasing process (let NLAR be their number) are chosen among the largest $R$ values having $|\Delta'| > \mathrm{SOG}$. The value of SOG is rather arbitrary and is not very critical: however, too small $|\Delta'|$ values might involve unreliable parameters in the triplet estimation process, too large values might extend the set of active reflections to magnitudes so small that, once phased, they would negligibly contribute to Fourier syntheses. In our tests, SOG and NLAR were chosen according to Table 2 where, for each structure, other useful parameters are specified: (*a*) the resolution of measured data for the derivative (RES); (*b*) the corresponding number of symmetry independent reflections (NREFL), which are simultaneously measured for native and for derivative; (*c*) the minimum value of $R$ among the NLAR reflections ($R_{\min}$); (*d*) the number of triplet invariants actively used in the phasing process (NTRIP). Triplets are estimated according to the reliability parameter [equation (11) of paper I]: the number of triplets estimated positive and negative by this equation are denoted in Table 2 by NPOS and NNEG, respectively.

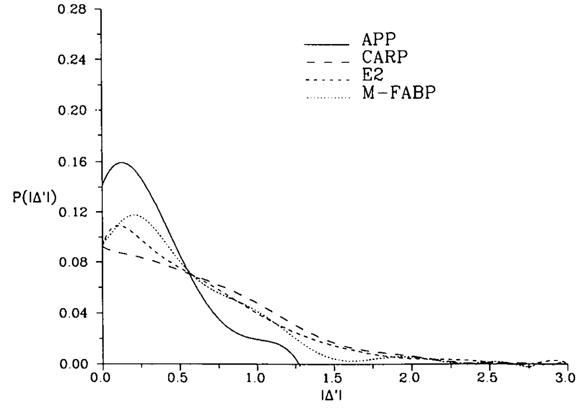

Fig. 3. Distribution of rescaled $|\Delta'|$'s for experimental data. The distribution refers to the NLAR reflections.

Before starting the phasing process, the program calculates the $z$ distribution (see paper I) relative to selected NLAR reflections and to NTRIP triplets stored for active use. The distributions for the four test structures (experimental data) are shown in Fig. 4: for all of them the statistical solvability criterion is satisfied and this gives the reasonable hope that all the structures can be solved. Seemingly, the worst situation occurs for APP (owing to the small value of NLAR), the most favourable occurs for CARP. The result for APP is mostly influenced by the small NLAR value and consequently by the reduced value of NTRIP. If for some structure the statistical solvability criterion is not verified, the user can suitably modify the values of SOG, NLAR and NTRIP. The CARP $z$ distribution starts at $z_{\min} \simeq 15$: it should be interpreted in terms of error effects rather than in terms of signal-to-noise ratio. The cumulative effect of the various sources of error makes the $|\Delta'|$'s much larger than they actually are (see the section later on dedicated to *post mortem* analysis of the phasing method).

The phasing procedure is a multisolution one, where starting sets of phases are generated by a random process (Baggio, Woolfson, Declercq & Germain, 1978). Random phases are given to NLAR/2 reflections (Burla, Cascarano & Giacovazzo, 1992) with unit weights for the origin and enantiomorph-fixing reflections and with weights equal to 0.8 for the others. Cycles of weighted tangent refinement are first applied to the NLAR/2 reflections and, after convergence, the phasing process is extended to all the NLAR reflections. As in *SIR*88 (Burla, Camalli, Cascarano, Giacovazzo, Polidori, Spagna & Viterbo, 1989) and *SIR*92 (Altomare, Cascarano, Giacovazzo, Guagliardi, Burla, Polidori & Camalli, 1994), a weighted tangent formula is used for phase extension and refinement:

$$\tan\varphi_{\mathbf{h}} = \sum_j \beta_j \sin(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j}) / \sum_j \beta_j \cos(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j})$$
$$= T_{\mathbf{h}}/B_{\mathbf{h}}, \qquad (2)$$

where $\beta_j$ is defined by the equation

$$D_1(\beta_j) = D_1(A)D_1(\alpha_{\mathbf{k}_j})D_1(\alpha_{\mathbf{h}-\mathbf{k}_j})$$

and

$$\alpha_{\mathbf{h}} = (T_{\mathbf{h}}^2 + B_{\mathbf{h}}^2)^{1/2}. \qquad (3)$$

The reliability parameter $\alpha_{\mathbf{h}}$ of any determined phase $\varphi_{\mathbf{h}}$ is modified according to the agreement between the calculated and the expected value of $\alpha_{\mathbf{h}}$. In particular, if $\alpha_{\mathbf{h}}$ is larger than the expected value

$$\langle\alpha_{\mathbf{h}}\rangle = \sum_i A_j D_1(A_j), \qquad (4)$$

Table 2. *Other useful parameters for the four structures*

DERIVATIVE denotes the atomic species added to the protein, RES [$=\lambda/(2\sin\theta_{\max})$] is the resolution of the measured data for the derivative.

| Structure code | DERIVATIVE | RES (Å) | NREFL | NLAR | SOG | $R_{\min}$ | NTRIP<br>(NPOS,NNEG) |
|---|---|---|---|---|---|---|---|
| APP | Hg | 2.0 | 2086 | 600 | 0.4 | 0.45 | 27437<br>(16459, 10978) |
| CARP | Hg | 2.0 | 4416 | 1000 | 0.8 | 0.7 | 30000<br>(13880, 16120) |
| E2 | Hg | 3.0 | 7757 | 1000 | 0.8 | 0.9 | 30000<br>(14946, 15054) |
| M-FABP | Hg | 3.0 | 2831 | 800 | 0.5 | 0.6 | 30000<br>(15232, 14768) |

then the calculated $\alpha_{\mathbf{h}}$ is replaced by

$$\langle\alpha_{\mathbf{h}}\rangle \exp[-(\alpha_{\mathbf{h}}-\langle\alpha_{\mathbf{h}}\rangle^2)/2\sigma^2_{\alpha_{\mathbf{h}}}]^{1/3},$$

where

$$\sigma^2_{\alpha_{\mathbf{h}}}=\tfrac{1}{2}\sum_j A_j^2[1+D_2(A_j)-2D_1^2(A_j)]. \qquad (5)$$

The weighting scheme is designed to drive phases towards values that minimize the difference between $\alpha$ and $\langle\alpha\rangle$ by reducing in the tangent refinement the importance of the phases with too large values of $\alpha$.

## Figures of merit

Recognizing the correct solution among different trials is not a simple task for protein structures (Woolfson & Yao, 1990; Giacovazzo, Guagliardi, Ravelli & Siliqi, 1994).

Figures of merit (FOMs) used in our procedure for picking the correct solution from the trial solutions are based on the theory described in two recent papers (Cascarano, Giacovazzo & Viterbo, 1987; Cascarano, Giacovazzo & Guagliardi, 1992*b*). Substantial modifications are, however, necessary to face the larger complexity of the problem and to take advantage of the information contained in derivative data.

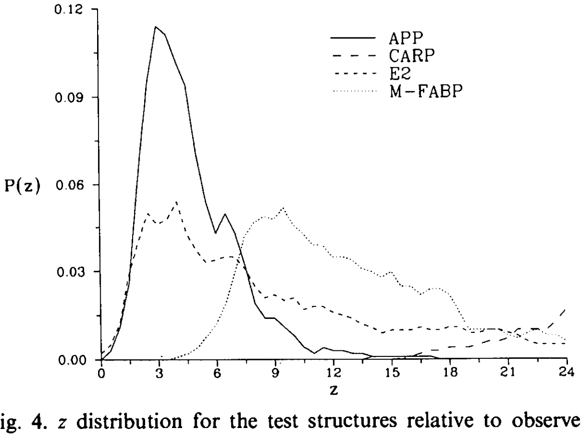

Fig. 4. *z* distribution for the test structures relative to observed data.

The first FOM is MABS $=(\sum_{\mathbf{h}}\alpha_{\mathbf{h}})/\langle\sum_{\mathbf{h}}\alpha_{\mathbf{h}}\rangle$, where

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

and

$$A_j=2[\sigma_3/\sigma_2^{3/2}]_p R_{\mathbf{h}}R_{\mathbf{k}_j}R_{\mathbf{h}-\mathbf{k}_j}+2[\sigma_3/\sigma_2^{3/2}]_H \Delta_{\mathbf{h}}\Delta_{\mathbf{k}_j}\Delta_{\mathbf{k}-\mathbf{k}_j}.$$

MABS gives a measure of the consistency of the triplet estimates but it is not used as an active FOM for picking (in combination with others) the correct solution.

The second FOM (ALFCOMB) depends on the ratios $(\alpha_{\mathbf{h}}-\langle\alpha_{\mathbf{h}}\rangle)/\sigma_{\alpha_{\mathbf{h}}}$, where $\sigma_{\alpha_{\mathbf{h}}}$ is given by (5).

This expression for the variance holds in the absence of errors in measurements and in their mathematical treatment as well as in the presence of perfect isomorphism between native and derivative structures. If this is not the case, as for real data, the variance cannot be perfectly calculated and is probably underestimated by $\sigma_{\alpha_{\mathbf{h}}}$. Accordingly, we used $2\sigma_{\alpha_{\mathbf{h}}}$ instead of $\sigma_{\alpha_{\mathbf{h}}}$ in ALFCOMB.

The third FOM (PSICOMB) relies on the expectation that the distribution of the psi-zero triplets should be as random as possible. PSICOMB depends on the ratios $\alpha'_{\mathbf{h}}/\sigma_{\alpha_{\mathbf{h}}'}$, where

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

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

$$\sigma_{\alpha_{\mathbf{h}}'}=\left(\sum_j A_j'^2\right)^{1/2}.$$

The weak reflections that constitute psi-zero triplets with the NLAR reflections are characterized by small values of both $R$ and $|\Delta'|$. Here, there is no room for a FOM based on classical negative quartet estimates based on native data only, which is unreliable for macromolecular structures of usual size.

Table 3. *APP*: *FOM values for the 'best' trial solutions as ranked by* CFOM

The last two lines correspond to the published refined structure (last line) and to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| 3 | 3.14 | 0.98 | 1.00 | 1.0 | 0.99 | 85 (84) |
| **2** | **2.75** | **0.94** | **1.00** | **1.0** | **0.96** | **41 (37)** |
| 24 | 1.72 | 0.47 | 1.00 | 1.0 | 0.65 | 85 (84) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 2.74 | 0.93 | 1.00 | 1.0 | 0.96 | 41 (37) |
|  | 1.41 | 0.92 | 0.99 | 0.94 | 0.93 | 0 (0) |

Table 4. *CARP*: *FOM values for the 'best' trial solutions as ranked by* CFOM

In the last line, FOM values corresponding to the true phase solution are quoted; the last but one line refers to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| 2 | 1.15 | 1.0 | 0.06 | 0.99 | 0.99 | 86 (86) |
| **21** | **0.93** | **0.0** | **0.34** | **0.87** | **0.42** | **41 (36)** |
| 3 | 0.70 | 0.0 | 0.02 | 0.70 | 0.33 | 84 (85) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 0.93 | 0.0 | 0.34 | 0.87 | 0.42 | 42 (36) |
|  | 0.50 | 0.0 | 0.12 | 0.45 | 0.22 | 0 (0) |

Table 5. *E2*: *FOM values for the 'best' trial solutions as ranked by* CFOM

In the last line, FOM values corresponding to the true phase solution are quoted; the last but one line refers to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| 22 | 1.59 | 1.0 | 0.95 | 1.0 | 1.0 | 87 (85) |
| 4 | 1.52 | 1.0 | 0.93 | 1.0 | 1.0 | 87 (88) |
| **10** | **0.94** | **0.24** | **0.88** | **0.88** | **0.52** | **30 (23)** |
| 9 | 0.56 | 0.02 | 0.35 | 0.60 | 0.26 | 86 (89) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 0.94 | 0.24 | 0.88 | 0.88 | 0.52 | 30 (23) |
|  | 0.70 | 0.15 | 0.60 | 0.63 | 0.36 | 0 (0) |

Table 6. *M-FABP*: *FOM values for the 'best' trial solutions as ranked by* CFOM

In the last line, FOM values corresponding to the true phase solution are quoted; the last but one line refers to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| **10** | **1.04** | **0.11** | **0.93** | **0.94** | **0.46** | **45 (37)** |
| 13 | 0.83 | 0.02 | 0.68 | 0.80 | 0.35 | 64 (61) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 1.04 | 0.11 | 0.94 | 0.94 | 0.46 | 44 (37) |
|  | 0.54 | 0.11 | 0.32 | 0.48 | 0.26 | 0 (0) |

In *SIR*88 and *SIR*92, a specific FOM (CPHASE) is based on negative estimated triplets by application of the $P_{10}$ formula (Cascarano, Giacovazzo, Camalli, Spagna, Burla, Nunzi & Polidori, 1984). In this context, negative and positive triplets here play a similar role: they are nearly equal in number and reliability and are both actively used in the phasing process. We therefore prefer to calculate the ratio

$$\sum_j A_j \cos \Phi_j / \sum_j A_j \langle \cos \Phi_j \rangle$$

for both positive and negative estimated triplet phases $\Phi_j$. A combined figure of merit (CFOM) integrates the indications arising from ALFCOMB, PSICOMB and CPHASE.

## Applications

The procedure described above has been applied to APP, CARP, E2 and M-FABP. The number of trials for each structure was fixed at 25, an amazing small number if one considers the complexity of the problem to be solved. FOM efficiency can be judged from Tables 3–6, where the trial solutions are ranked in decreasing order with respect to the combined figure of merit CFOM. Each FOM must lie between zero and one and is expected to be one for the correct solution. On the same line, the trial number and the values of MABS, ALFCOMB, PSICOMB, CPHASE and CFOM are shown. In the last column, the average phase error $\langle|\varphi_{\rm true} - \varphi_{\rm calc}|\rangle$ (unweighted and weighted) is given in degrees (ERR). The last two lines show the same data for the solution that is obtained by tangent refinement of the true phases and for the published refined structure (last line). The information contained in the last line suggests to the reader some inefficiency of the FOMs: often they are not maximal for the correct structure. The last line provides a guess of the overall reliability of triplet estimates used in the tangent refinement process: good triplets usually lead to small values of ERR.

Tables 3–6 show that FOMs are not optimal but are sufficiently good for practical purposes. The order of the correct solution (bold typeface) is 2 for APP and CARP, 1 for M-FABP and 3 for E2. The weighted error is not larger than 37° for all the test structures: the best performance occurs for E2 (23°); this has the largest unit cell but also an unusually good derivative. It is worthwhile noting that the correct solution found among the 25 trials has an average phase error very close to the solution obtained by tangent refinement of the true phases. In other words, the procedure easily drives phases to converge to the 'best' values allowed by the overall efficiency of triplet relationships.

It is not possible from Tables 3–6 to derive conclusions about the relative efficiencies of the various FOM's. For example: (*a*) PSICOMB is highly discriminant for CARP and M-FABP but absolutely useless for APP; (*b*) ALFCOMB is discriminant for M-FABP and absolutely useless for CARP; CPHASE works quite well for M-FABP and is of modest discriminating power for the others. This oscillating behaviour is better explained in the section dedicated to the *post mortem* analysis of the phasing method. A general feeling of the quality of the Fourier maps obtained by our procedure may be gathered from Figs. 5 and 6, where portions of the M-FABP electron-density map are shown. In Fig. 5(*a*), the electron-density map around residues 19–22 (Tyr-Met-Lys-Ser) is shown as obtained at the end of our phasing process (800 phased reflections). The figure can usefully be compared with the refined electron-density map calculated by Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992) at the end of their structure-refinement process by using all reflections to 2.1 Å resolution (Fig. 5*b*). The correlation between the two maps is remarkable. Differences are due mostly to the limited number of phased reflections used for Fig. 5(*a*) rather than to phase errors. This is confirmed by Fig. 5(*c*), where correct phases are associated with the 800 reflections used for Fig. 5(*a*).

Of particular interest is Fig. 6(*a*), where electron density, as obtained by our procedure for a portion of $\alpha$-helix II (residues from 27 to 36), is shown. Figs. 6(*b*) and (*c*) represent the electron density calculated for the same region as described for Figs. 5(*b*) and

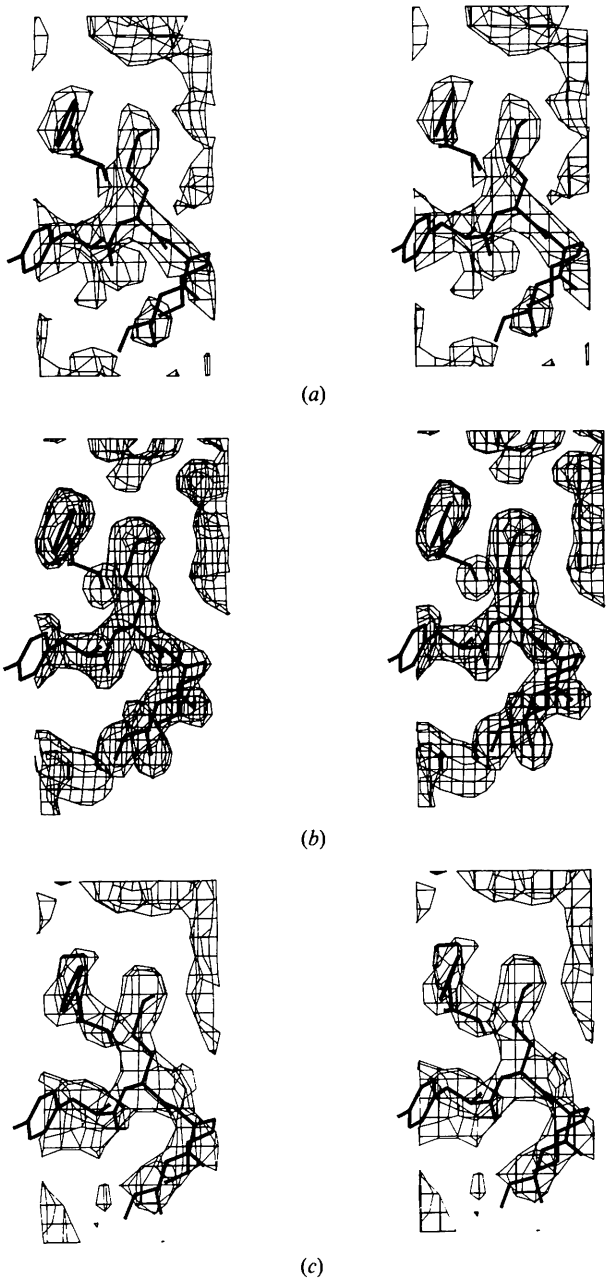

Fig. 5. Electron-density maps around residues 19–22 (Tyr-Met-Lys-Ser) of M-FABP, calculated with: (*a*) 800 reflections, phased from direct methods; (*b*) all reflections to 2.1 Å resolution with true phases; (*c*) the same 800 reflections as in (*a*) associated with true phases (stereo).

(*c*). Maps 6(*a*) and (*c*) show similar breaks in the electron density: these are in correspondence with nitrogen 32 and with peptide bonds 29–30 and 35–36.

Figs. 5 and 6 are examples of good correlation between calculated and true electron dẹnsities. Unfortunately, examples of bad correlation can frequently be found.

Generally, the electron-density maps we obtain suffer from the same drawbacks usually found for small molecules. For example, owing to series-termination errors, some structural parts are well emphasized with respect to the background, some others drown in it and false details are generated. Continuity in the electron density, an important requisite for macromolecules (which would make chemical interpretation of the Fourier map easier), is not guaranteed at this stage of the phasing process and the map is therefore not interpretable. The general conclusion is that phase extension rather than a better phase refinement is the most urgent problem to be faced.

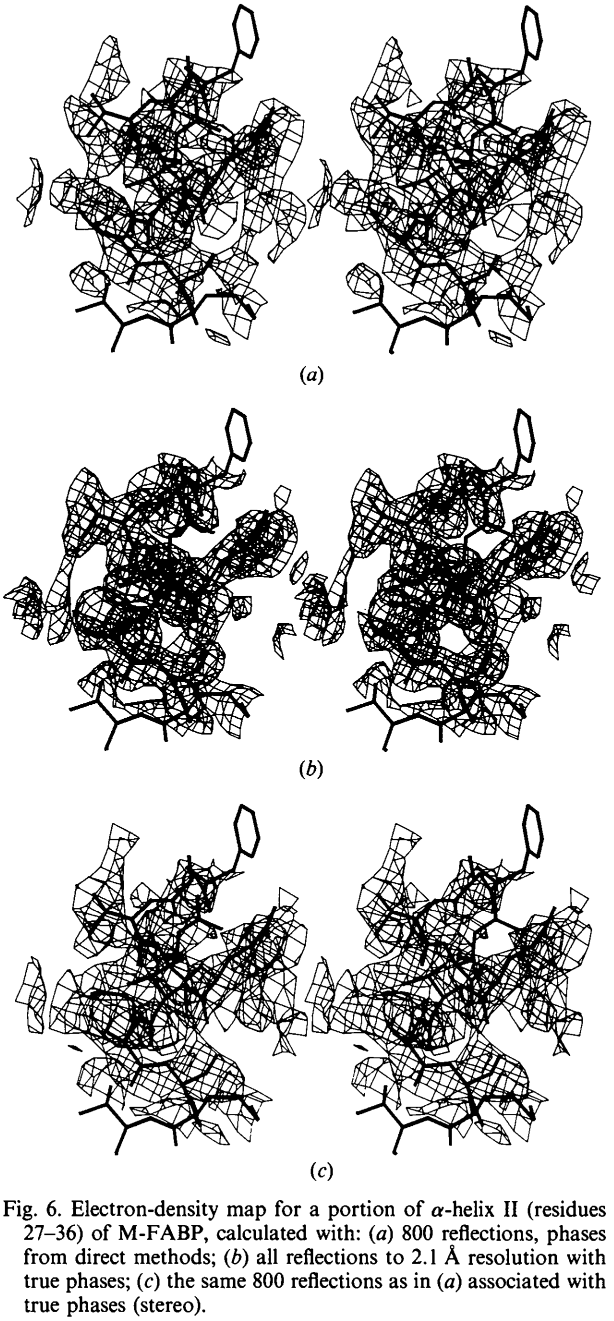

Fig. 6. Electron-density map for a portion of $\alpha$-helix II (residues 27–36) of M-FABP, calculated with: (*a*) 800 reflections, phases from direct methods; (*b*) all reflections to 2.1 Å resolution with true phases; (*c*) the same 800 reflections as in (*a*) associated with true phases (stereo).

### Loss of enantiomorph

Any direct-methods user working in the small-molecule field knows perfectly well that a perverse property of the tangent formula is the tendency to lose the enantiomorph in the 73 symmorphic and in several nonsymmorphic space groups. The tendency can be fought by introducing special weighting schemes in the tangent formula and by designing FOMs that are able to discard false centrosymmetric or pseudocentrosymmetric solutions from the set of the most probable trials. The space groups of APP, CARP and M-FABP belong to the subset for which the enantiomorph is easily lost: actually, we obtain for APP and CARP pseudocentrosymmetric solutions associated with high FOM values. For example, in APP the correct position of the Zn atom is clearly and correctly defined but a pseudocentrosymmetric image is also obtained. This is all in spite of the fact that our weighting scheme was designed to hinder the loss of enantiomorph and our FOMs were devoted to discarding centrosymmetric solutions. Any attempt at modifying the weighting scheme for preserving the enantiomorph failed. The question becomes clear if one looks at Tables 7–8: by using error-free data we partitioned the NREFL reflections of APP and CARP into subsets, each subset including $|\Delta'|$ values falling in a fixed interval. The tables show for APP and CARP the number of reflections (nr) per subset and the average phase difference between the true phase and $(0,\pi)$. Reflections with very small $|\Delta|$ values are more probably close to $\pm\pi/2$ than to $n\pi$, while reflections with large $|\Delta|$ values are pseudocentrosymmetric. The obvious conclusion is that the conditions adopted for the selection of the NLAR reflections (*i.e.* $R$ and $|\Delta'|$ sufficiently large) defines at the same time a subset of phases that are in a large majority centrosymmetric. Enantiomorph-sensitive phases are not present in the group of NLAR reflections: that is the reason why the enantiomorph *cannot* be maintained for CARP and APP in the framework of the present method. We need a supplementary step where enantiomorph-sensitive phases would be involved and play a central

Table 7. *APP: the number of reflections* (nr) *with* $|\Delta'|$ *lying in a given interval and the corresponding average phase difference* ($\langle|\Delta\Phi|\rangle$) *between true phases and* $(0,\tau)$

Error-free calculated data are used.

| Range | nr | $\langle\|\Delta\Phi\|\rangle$ |
|---|---|---|
| 0.0–0.1 | 302 | 52 |
| 0.1–0.2 | 323 | 48 |
| 0.2–0.3 | 295 | 48 |
| 0.3–0.4 | 225 | 45 |
| 0.4–0.5 | 198 | 41 |
| 0.5–0.6 | 174 | 35 |
| 0.6–0.7 | 158 | 36 |
| 0.7–0.8 | 106 | 29 |
| 0.8–0.9 | 100 | 27 |
| 0.9–1.0 | 85 | 21 |
| 1.0–1.1 | 69 | 13 |
| 1.1–1.2 | 28 | 7 |
| 1.2–1.3 | 14 | 4 |
| 1.3–1.4 | 8 | 2 |
| 1.4–1.5 | 1 | 11 |
| 1.5–1.6 | 1 | 0 |

Table 8. *CARP: the number of reflections* (nr) *with* $|\Delta'|$ *lying in a given interval and the corresponding average phase difference* ($\langle|\Delta\Phi|\rangle$) *between true phases and* $(0,\pi)$

Error-free calculated data are used.

| Range | nr | $\langle\|\Delta\Phi\|\rangle$ |
|---|---|---|
| 0.0–0.1 | 838 | 62 |
| 0.1–0.2 | 702 | 57 |
| 0.2–0.3 | 617 | 50 |
| 0.3–0.4 | 506 | 46 |
| 0.4–0.5 | 438 | 38 |
| 0.5–0.6 | 438 | 35 |
| 0.6–0.7 | 397 | 27 |
| 0.7–0.8 | 297 | 21 |
| 0.8–0.9 | 311 | 15 |
| 0.9–1.0 | 112 | 8 |
| 1.0–1.1 | 29 | 4 |
| 1.1–1.2 | 1 | 0 |

role. This concerns the phase-extension process rather than the present paper. It cannot, however, be claimed that the enantiomorph is always lost by our procedure in symmorphic space groups. E2 is just one example of a structure crystallizing in a symmorphic space group for which the enantiomorph is not lost. The problem for CARP is probably magnified by the fact that distribution of calculated phase angles for the native protein is essentially flat between 30 and 150° and peaks sharply in the $-2.5$ to 2.5° interval. Of the 4829 noncentrosymmetric reflections, there are 733 more near 0 or 180° than predicted by a random distribution of relative phase angles (Kretsinger & Nockolds, 1973).

## *Post mortem* analysis of the phasing method

The limits and potential of our phasing method cannot be fully understood without a *post mortem* analysis of the efficiency of the various steps of our procedure. It will than be possible to recognize the most critical points of the process and design better strategies for a more robust phasing process.

First, we consider the normalization step. A guess about errors introduced by the differential Wilson plot can be obtained by using ideal error-free data and by calculating the percentage of reflections that undergo sign inversion for $\Delta'$ (*i.e.* the true sign of $\Delta'$, calculated from known native and derivative structures, changes as a result of the errors introduced by the normalization process). The variation of this percentage as a function of $|\Delta'|$ is shown in Fig. 7 for the four test structures (curves $P_{\rm inv}$) and with respect to the NREFL reflections. The inversion rate due to the statistical treatment of data is not negligible for, say, $|\Delta'| < 0.2$.

Let us now repeat the same calculations using experimental data. This time, the sign inversion for $\Delta'$ is due to the combined action of physical sources of error (mostly lack of isomorphism and errors in measurements) and of their statistical treatment in the normalization process. The curves $P_{\rm inv}$ for the four test structures are shown in Fig. 8. The percentage of sign inversions is now much larger than for calculated data: surprisingly, $P_{\rm inv}$ is not negligible even for very large $|\Delta'|$ values. APP and E2 have the most favourable behaviours: $P_{\rm inv}$ is about 0.10 for $|\Delta'| \simeq 1$ but increases to 0.20 for M-FABP. $P_{\rm inv}$ is amazingly high for CARP: it is nearly constant (about 0.45) over a large range and even increases for the largest $|\Delta'|$'s. It is instructive to compare in Table 9 the 51 observed $|\Delta'| > 2.0$ found among the NREFL reflections, with the corresponding true (error-free) calculated values. Among several large errors there are 'impossible' observed $\Delta'$ values: *e.g.* 5.23, 6.21 and 7.74, for which the calculated values are 0.0, $-0.07$ and $-0.12$. A large $|\Delta'|$ value for CARP is by no means a warranty that its sign is correct. No correlation is found between the error

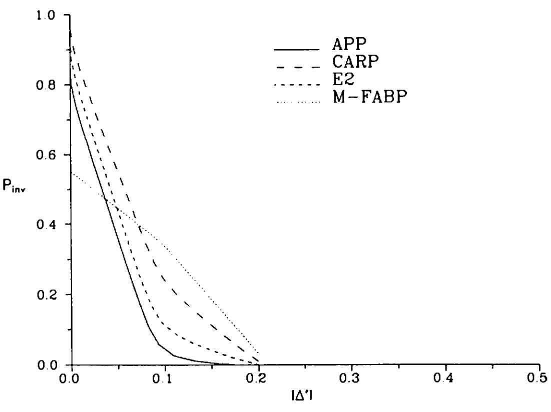

Fig. 7. Percentage of reflections that undergo sign inversion for $\Delta'$ as a result of the normalization process (calculated error-free data).

magnitude and the |E| value. Table 9 corroborates our decision to fix at 2 the $|\Delta'|$'s larger than 2.

Since a sign inversion for $\Delta'_{\mathbf{h}}$ changes the expected values of all the triplets in which **h** is involved, one can wonder how is it possible to solve the CARP structure when about 0.43 of the $\Delta$ signs are wrong. We do not have a final answer to the problem: perhaps it depends on a possible organized mechanism for sign inversion. However, a final conclusion can certainly be drawn: since all our test structures, CARP included, are solved by our procedure, the common belief that direct methods are too sensitive to experimental (*latu sensu*) errors, and therefore the feeling that they are unable to solve *ab initio* protein structures, must be rejected. Our procedure proved extremely robust even for CARP. The key to our success is the simultaneous use of many phase relationships: even if single ones are heavily affected by errors, all of them cooperate against failure.

Let us now consider the quality of the prime tools of the phasing process, the triplets. The larger the percentage of reflections showing sign inversion for $\Delta'$, the lower will be the efficiency of the reliability parameter, equation (11) of paper I, for triplet estimation. In Table 10, we show for each test structure some statistical calculations performed on triplets really employed (experimental data) in the phasing process. In the table, Nr is the number of triplets having $|A| > |\mathrm{ARG}|$, % is the percentage of triplets whose cosine sign is correctly estimated, $\langle|\Phi|\rangle$ is the average of the absolute values of the triplet phase $\Phi$. Comparison with Table 6 of paper I relative to free calculated data shows that the efficiency of the parameter given by equation (11) of paper I drops as a consequence of the experimental errors in the data, lack of isomorphism, statistical treatment of the data *etc.* As could be expected, the worst situation occurs for CARP, where, as an effect of the 'experimental' errors, the sign inversion for $\Delta'$ is particularly frequent. The percentage of correctly estimated triplets is small and uniformly distributed. Furthermore, too high reliability parameters $|A|$ are obtained in correspondence with high percentages of wrong triplets.

Table 9. *The* 51 *observed* $\Delta'$ *values* (*among the* NREFL *reflections*) *with* $|\Delta'| \geq 2.0$ *and the corresponding calculated* (*error-free*) *values*

| $\lvert E\rvert$ | $\Delta'_{\rm obs}$ | $\Delta'_{\rm calc}$ | $\lvert E\rvert$ | $\Delta'_{\rm obs}$ | $\Delta'_{\rm calc}$ |
|---|---|---|---|---|---|
| 2.61 | −2.16 | −0.66 | 1.08 | −2.05 | −0.65 |
| 2.24 | −3.50 | 0.80 | 1.06 | 2.54 | −0.78 |
| 2.08 | −2.14 | 0.57 | 1.03 | −2.04 | −0.39 |
| 2.07 | −2.00 | −0.08 | 1.03 | −2.03 | −0.66 |
| 1.98 | −2.54 | −0.72 | 1.02 | −2.29 | 0.42 |
| 1.93 | −2.30 | −0.63 | 0.99 | 3.04 | −0.03 |
| 1.88 | −2.13 | −0.69 | 0.97 | −2.21 | 0.50 |
| 1.88 | −2.07 | −0.57 | 0.93 | 2.27 | 0.15 |
| 1.82 | 3.62 | −0.09 | 0.93 | −2.02 | 0.40 |
| 1.71 | −2.60 | −0.99 | 0.92 | 2.24 | −0.72 |
| 1.64 | −3.68 | 0.24 | 0.92 | −2.10 | −0.82 |
| 1.51 | 5.23 | 0.00 | 0.91 | 2.38 | 0.44 |
| 1.51 | −3.08 | −0.66 | 0.91 | −2.02 | 0.32 |
| 1.48 | −2.30 | 0.62 | 0.83 | 3.81 | −0.92 |
| 1.41 | −2.31 | 0.01 | 0.76 | 2.19 | 0.50 |
| 1.40 | −2.03 | 0.19 | 0.75 | 3.50 | 0.64 |
| 1.38 | −2.03 | −0.40 | 0.73 | 2.00 | −0.15 |
| 1.26 | −2.01 | −0.02 | 0.71 | 3.61 | 0.77 |
| 1.25 | −2.23 | 0.00 | 0.68 | 2.10 | −0.27 |
| 1.24 | −2.18 | −0.99 | 0.60 | 2.31 | 0.87 |
| 1.23 | −2.56 | 0.41 | 0.60 | 2.44 | 0.16 |
| 1.21 | −2.36 | 0.09 | 0.59 | 4.28 | 0.82 |
| 1.21 | −2.18 | 0.31 | 0.52 | 6.21 | −0.07 |
| 1.16 | −2.68 | 0.48 | 0.32 | 7.74 | −0.12 |
| 1.13 | 2.47 | 0.17 | 0.26 | 2.72 | 0.45 |
| 1.12 | −2.44 | 0.22 | | | |

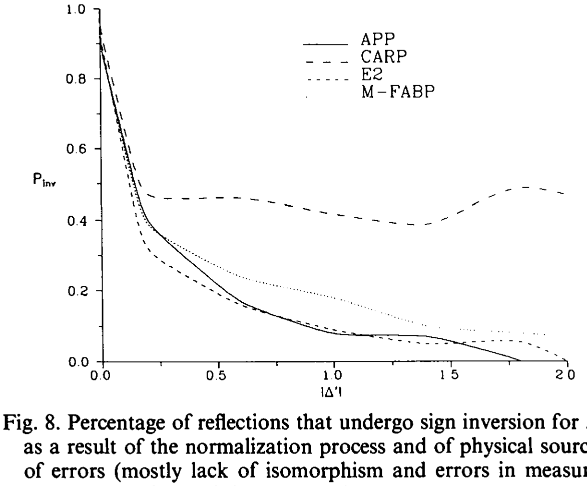

Fig. 8. Percentage of reflections that undergo sign inversion for $\Delta'$ as a result of the normalization process and of physical sources of errors (mostly lack of isomorphism and errors in measurements). Experimental data are used.

One should wonder why the relatively small percentage of triplets correctly estimated is able to drive phases to nearly correct values. The question is not of minor importance if one observes that among the 'correctly estimated triplets' we also include triplets having estimated phase substantially differing from the true value (sometime up to 90° of difference). In our opinion, the reserve of power of the method lies in the existence of a nearly equivalent number of estimated positive and negative triplets. Errors balance each other out and a solution can be attained in a rather straightforward way. In these conditions, the accuracy of the final phases should be modest but phase values are useful and fix the structure. A countercheck for this conclusion may be obtained by using in the phasing process error-free calculated data: in these conditions, accurate phase values should be produced by the phasing process. The results are shown in Table 11. The order of the correct solution (among the various trials) as ranked by CFOM is always 1 (*i.e.* that with the highest value of CFOM) for all the test structures and the final mean phase error calculated over the NLAR reflections is comparable with corresponding errors for small molecules.

The strange behaviour of the figures of merit can be explained by comparing Tables 3–6, obtained for

Table 10. *Statistical calculations for triplet invariants estimated via equation* (11) *of paper* I *for various values of* SOG *used in the phasing process*

Measured data for native and derivative structures are used.

| | Positive estimated triplets | | | Negative estimated triplets | | |
|---|---|---|---|---|---|---|
| $\lvert ARG \rvert$ | Nr | % | $\langle \lvert \Phi \rvert \rangle$ | Nr | % | $\langle \lvert \Phi \rvert \rangle$ |
| APP (SOG = 0.4) | | | | | | |
| 0.2 | 16459 | 69.7 | 67.5 | 10978 | 67.1 | 109.7 |
| 0.4 | 12499 | 71.7 | 65.4 | 4541 | 71.4 | 114.7 |
| 0.8 | 3104 | 77.7 | 58.3 | 824 | 75.0 | 119.0 |
| 1.2 | 831 | 81.3 | 53.0 | 196 | 74.0 | 119.5 |
| 1.6 | 256 | 88.3 | 45.0 | 36 | 72.2 | 124.6 |
| 2.0 | 83 | 90.4 | 43.9 | 10 | 80.0 | 135.5 |
| 2.6 | 14 | 85.7 | 51.6 | 2 | 50.0 | 132.0 |
| 3.2 | 2 | 100.0 | 69.0 | | | |
| CARP (SOG = 0.8) | | | | | | |
| 2.6 | 13880 | 72.0 | 64.2 | 16120 | 65.3 | 107.6 |
| 3.2 | 10452 | 72.6 | 63.4 | 11442 | 65.6 | 108.0 |
| 4.4 | 4748 | 73.6 | 62.3 | 5347 | 67.6 | 110.2 |
| 6.5 | 1180 | 73.7 | 61.9 | 1415 | 70.1 | 112.8 |
| 9.0 | 226 | 70.8 | 64.8 | 295 | 73.6 | 117.5 |
| 15.0 | 4 | 75.0 | 73.2 | 4 | 50.0 | 110.2 |
| E2 (SOG = 0.8) | | | | | | |
| 0.8 | 14946 | 73.8 | 62.2 | 15054 | 71.7 | 115.5 |
| 1.6 | 6143 | 77.6 | 57.9 | 5457 | 75.0 | 119.6 |
| 2.6 | 381 | 86.6 | 46.6 | 340 | 84.7 | 129.1 |
| 3.2 | 27 | 81.5 | 51.3 | 34 | 67.6 | 110.7 |
| M-FABP (SOG = 0.5) | | | | | | |
| 0.8 | 15232 | 64.2 | 73.8 | 14768 | 61.7 | 103.8 |
| 1.6 | 4596 | 68.6 | 68.9 | 3696 | 69.0 | 111.4 |
| 2.6 | 1002 | 74.0 | 63.4 | 742 | 73.2 | 117.0 |
| 3.8 | 214 | 76.6 | 58.9 | 172 | 77.3 | 124.0 |
| 5.5 | 31 | 80.6 | 54.9 | 23 | 87.0 | 132.7 |
| 6.5 | 7 | 100.0 | 35.1 | 8 | 87.5 | 142.7 |

Table 11. *Mean phase error at the end of the phasing process relative to the* NLAR *reflections*

Error-free calculated data are used.

| Structure code | $\langle \lvert \Phi_{\text{true}} - \Phi_{\text{calc}} \rvert \rangle$ (°) | $\langle w \lvert \Phi_{\text{true}} - \Phi_{\text{calc}} \rvert \rangle$ (°) | Order of solution |
|---|---|---|---|
| APP | 26 | 25 | 1 |
| CARP | 20 | 19 | 1 |
| E2 | 15 | 17 | 1 |
| M-FABP | 20 | 20 | 1 |

experimental data, with Tables 12–15 obtained for error-free calculated data. For example, when experimental data are used, ALFCOMB is zero for the true solution of CARP, close to zero for the true solution of M-FABP and very small for the true solution of E2. In contrast, ALFCOMB is close to 1 for the true solution of all the test structures when calculated data are used. Again, PSICOMB is very small for the correct solution of CARP when experimental data are used and close to 1 when calculated data are used. The examples clearly indicate that large 'errors' connected to experimental data can greatly disturb the efficiency of the various FOMs in a rather unforeseeable way. The search for FOMs less sensitive to 'experimental' errors is a topic of enormous importance; however, the FOMs proposed in this paper can be considered a first important step in this direction.

A last observation concerns the extraordinary small number of trials necessary for obtaining a correct solution when error-free data are used. Three trials are sufficient for all the test structures. This surprising result, not attainable even for small molecules, suggests that direct phase solution will be even easier when the various sources of error are depressed by new experimental techniques and/or by a more efficient mathematical treatment.

## Concluding remarks

A direct procedure for *ab initio* crystal structure solution of proteins has been described. The method is based on a probabilistic approach that integrates direct methods and isomorphous techniques. The keystone is the formula estimating three-phase invariants given six magnitudes obtained by Giacovazzo, Cascarano & Zheng (1988), in which the crucial role of the normalized difference $\Delta'$ is emphasized. The combined use of native and derivative data imposes remarkable differences between our procedure and the typical ones used for small-molecule crystal structure solution. In order to emphasize the necessary differences, the procedure has been described step by step: for each step, reasons for the various choices are presented.

The procedure seems to be efficient and robust: the correct solution is obtained after a few trials and is ranked in the first positions by suitable figures of merit. However, we do not claim that it is optimal: in spite of its great success (for the first time, the successful solution of the phase problem *ab initio* by direct methods for a non-negligible set of reflections is performed), the procedure still has some drawbacks (*i.e.* the possible loss of the enantiomorph in some symmorphic space groups and the relatively small number of phased reflections) but has also a reserve of power. Indeed, each of its steps may be remarkably improved: the normalization process, the weighting scheme used in the tangent formula for phase extension and refinement, the treatment of the experimental errors, the reduction of the effects of lack of isomorphism and the figures of merit are all areas that may benefit from future contributions. In particular, those methods that analyse the effect of measurement errors in the routine methods of isomorphous replacement could profitably be used for the supplementary weighting of estimated triplet phases. In conclusion, the potential of the procedure here described is far from being exhausted. Two trivial examples of the limitations of the present procedure are as follows.

(*a*) To spare computer central memory, the set of active triplets has been limited to 30000. In this way,

Table 12. *APP: calculated error-free data; FOM values for the 'best' trial solutions as ranked by* CFOM

In the last line, FOM values relative to the true phase solution are quoted; the last but one line refers to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| **3** | **2.21** | **0.96** | **1.0** | **1.0** | **0.97** | **26 (26)** |
| 1 | 2.23 | 0.95 | 1.0 | 1.0 | 0.96 | 86 (86) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 2.22 | 0.96 | 1.0 | 1.0 | 0.97 | 26 (26) |
|  | 1.61 | 0.92 | 1.0 | 1.0 | 0.95 | 0 (0) |

Table 13. *CARP: calculated error-free data; FOM values for the 'best' trial solutions as ranked by* CFOM

In the last line, FOM values relative to the true phase solution are quoted; the last but one line refers to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| **3** | **2.13** | **1.0** | **1.0** | **1.0** | **1.0** | **20 (19)** |
| 5 | 1.25 | 0.15 | 0.99 | 1.0 | 0.48 | 85 (84) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 2.13 | 1.0 | 1.0 | 1.0 | 1.0 | 20 (19) |
|  | 1.81 | 1.0 | 1.0 | 1.0 | 1.0 | 0 (0) |

Table 14. *E2: calculated error-free data; FOM values for the 'best' trial solutions as ranked by* CFOM

In the last line, FOM values relative to the true phase solution are quoted; the last but one line refers to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| **1** | **1.29** | **0.98** | **1.0** | **1.0** | **0.99** | **15 (17)** |
| **7** | **1.29** | **0.98** | **0.99** | **1.0** | **0.98** | **15 (17)** |
| 2 | 1.60 | 0.91 | 1.0 | 1.0 | 0.95 | 85 (82) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 1.29 | 0.98 | 1.0 | 1.0 | 0.99 | 15 (17) |
|  | 1.10 | 0.95 | 0.93 | 0.92 | 0.94 | 0 (0) |

Table 15. *M-FABP: calculated error-free data; FOM values for the 'best' trial solutions as ranked by* CFOM

In the last line, FOM values relative to the true phase solution are quoted; the last but one line refers to the solution obtained by tangent refinement of the true phases.

| Trial | MABS | ALFCOMB | PSICOMB | CPHASE | CFOM | ERR (weighted) |
|---|---|---|---|---|---|---|
| **2** | **2.32** | **0.93** | **1.0** | **1.0** | **0.96** | **19 (20)** |
| 4 | 1.43 | 0.22 | 1.0 | 1.0 | 0.50 | 48 (47) |
| 5 | 1.51 | 0.21 | 1.0 | 1.0 | 0.50 | 82 (87) |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
|  | 2.32 | 0.93 | 1.0 | 1.0 | 0.953 | 19 (20) |
|  | 1.86 | 0.91 | 1.0 | 1.0 | 0.94 | 0 (0) |

phase information contained in other triplets is lost: for example, for E2, 329 164 triplets are found among the 1000 active reflections, of which 163 632 are estimated positive and 165 532 negative. 299 164 of such triplets are presently neglected. (*b*) NLAR cannot be too large because of point (*a*). Less restrictions in the procedure will probably improve the efficiency of the phasing process. Furthermore, the method will strongly benefit from the most modern experimental techniques aimed at producing higher-quality crystals for native and derivative structures and at reducing measurement errors. In our opinion, even larger structural systems could in principle be accessible to this phasing process. The main limitation of the present approach is the small number of phased reflections rather than the quality of the assigned phases. As a consequence, continuity in the electron-density map is not secured and map interpretation (for example, the correct tracing of the whole polypeptide backbone) is not possible at this stage. However, the assigned phases are of such high quality that they can act as a starting point for a complete protein structure determination. Future efforts will be devoted to the extension of phases to reflections not involved in the present procedure. It will be shown in a subsequent paper that this can be performed without impoverishing the quality of the phase values.

The authors thank Professor Giuseppe Zanotti and Dr Andrea Mattevi for many stimulating discussions.

### References

Altomare, A., Cascarano, G., Giacovazzo, C., Guagliardi, A., Burla, M. C., Polidori, G. & Camalli, M. (1994). *J. Appl. Cryst.* **27**, 435.

Baggio, R., Woolfson, M. M., Declercq, J.-P. & Germain, G. (1978). *Acta Cryst.* **A34**, 883–892.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*, p. 336. London: Academic Press.

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P. & Wu, C. W. (1981). *Proc. Natl Acad. Sci. USA*, **7**, 4175–4179.

Burla, M. C., Camalli, M., Cascarano, G., Giacovazzo, C., Polidori, G., Spagna, R. & Viterbo, D. (1989). *J. Appl. Cryst.* **22**, 389–393.

Burla, M. C., Cascarano, G. & Giacovazzo, C. (1992). *Acta Cryst.* **A48**, 906–912.

Cascarano, G., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* **A40**, 278–293.

Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1992*a*). *Z. Kristallogr.* **200**, 63–71.

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

Cascarano, G., Giacovazzo, C. & Viterbo, D. (1987). *Acta Cryst.* **A43**, 22–29.

Giacovazzo, C., Casarano, G. & Zheng, C.-D. (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.

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

Hall, S. R. & Subramanian, V. (1982). *Acta Cryst.* **A38**, 590–598.

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

Kretsinger, R. H. & Nockolds, C. E. (1973). *J. Biol. Chem.* **248**, 3313–3326.

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

Woolfson, M. M. & Yao, J.-X, (1990). *Acta Cryst.* **A46**, 409–413.

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

*Acta Cryst.* (1994). **A50**, 621–632

# Thermal Vibrations and Bonding in GaAs: an Extended-Face Crystal Study

By Andrew W. Stevenson

*CSIRO Division of Materials Science and Technology, Private Bag* 33, *Rosebank MDC, Clayton,*
*Victoria* 3169, *Australia*

(*Received* 14 *June* 1993; *accepted* 21 *March* 1994)

## Abstract

Accurate X-ray integrated-intensity data collected from an extended-face crystal of GaAs are analysed to provide detailed information on the thermal vibrations of atomic species, including cubic anharmonicity, at room temperature. The values obtained for the thermal parameters are $B_{\rm Ga} = 0.622\,(3)$ Å$^2$, $B_{\rm As} = 0.483\,(5)$ Å$^2$ and $\beta_{\rm GaAs} = -0.6\,(1) \times 10^{-18}$ J Å$^{-3}$ (defined in the text). The inclusion of cubic anharmonic thermal vibrations is shown to be highly significant. In order to interpret the data collected for certain low-angle Bragg reflections for which $h+k+l=4n+2$ (in particular, 200, 222 and $22\bar{2}$), it is necessary to consider bonding effects. It is shown that there is a net transfer of electron charge from gallium to arsenic [$Q = 0.12\,(3)$ e] and that the inclusion of bonding effects in the least-squares analysis is highly significant. The analysis includes allowance for the extremely severe extinction effects present for such a perfect sample (minimum extinction factor 0.286). The refined value of the mean radius of perfect-crystal domains is 4.6 (2) μm. The final fit, for 153 independent Bragg reflections, is excellent, as indicated by the weighted $R$ factor of 0.683% and the goodness-of-fit parameter of 1.083. The results of the least-squares analysis are compared for the cases of relativistic Hartree–Fock, Thomas–Fermi–Dirac and relativistic Dirac–Slater atomic scattering factors, the former being favoured.

## Introduction

GaAs is an extremely important semiconductor material that possesses the sphalerite (zinc blende) structure. Knowledge of the way in which the atomic species in GaAs vibrate is important in many areas of research such as studies of diffusion and for predicting band-gap temperature dependence. A survey of the literature shows that reported values of individual Debye–Waller factors for gallium and arsenic in GaAs, both experimental and theoretical, show a large variation (see also Butt, Bashir & Nasir