the asymmetry and then the width of the rocking curve by a simple rotation of the sample; (ii) the use of very inclined reflecting planes in a tilted and symmetric geometry enables a decrease in the thermal load on monochromators, since the trace of the incident beam on the surface of the crystal is then much larger than in the case of symmetric reflections on the surface (Macrander *et al.*, 1992).

The experiments were performed on beam line D15B of the DCI storage ring at LURE, Orsay, France. We thank very much Professor J. Derrien of the CRMC2, Marseille, France, for the preparation of the sample. This work was financially supported by the French Ministry of Research and the EEC (Esprit BRA no. 3026).

References

AUTHIER, A. (1986). *Acta Cryst.* A**42**, 414–426.
BATTERMAN, B. W. (1964). *Phys. Rev. A*, **133**, 759–764.
BATTERMAN, B. W. (1969). *Phys. Rev. Lett.* **22**, 703–705.
BEDZYK, M. J., GIBSON, W. M. & GOLOVCHENKO, J. A. (1982). *J. Vac. Sci. Technol.* **20**, 634–637.
BEDZYK, M. J. & MATERLICK, G. (1985). *Surf. Sci.* **152/153**, 10–16.
BOULLIARD, J. C., CAPELLE, B., FERRET, D., LIFCHITZ, A., MALGRANGE, C., PÉTROFF, J. F., TACCOEN, A. & ZHENG, Y. L. (1992). *J. Phys.* I (France), **2**, 1215–1232.
MACRANDER, A. T., LEE, W. K., SMITH, R. K., MILLS, D. M., ROGERS, C. S. & KHOUNSARY, A. M. (1992). *Nucl. Instrum. Methods Phys. Res.* A**319**, 188–196.
SAITOH, Y., HASHIZUME, H. & TSUTSUI, K. (1988). *Jpn. J. Appl. Phys.* **27**, 1386–1396.
VLIEG, E., FONTES, E. & PATEL, J. R. (1991). *Phys. Rev. B*, **43**, 7185–7193.
ZEGENHAGEN, J. (1993). *Surf. Sci. Rep.* **18**, 199–271.

Acta Cryst. (1994). A**50**, 503–510

The *Ab Initio* Crystal Structure Solution of Proteins by Direct Methods. I. Feasibility

BY CARMELO GIACOVAZZO AND DRITAN SILIQI*

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

AND ADAM RALPH†

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

(*Received* 8 *October* 1993; *accepted* 4 *January* 1994)

Abstract

Traditional direct methods based on the tangent formula and/or on Sayre's equation cannot solve *ab initio* the large majority of protein crystal structures [Giacovazzo, Guagliardi, Ravelli & Siliqi (1994). *Z. Kristallogr.* **209**, 136–142]. Indeed, the amount of information available leads to a signal-to-noise ratio close to unity; consequently, the correct solution, even if attained, cannot be recognized among the trial solutions. Attention is here focused onto the case in which diffraction data of one isomorphous derivative are additionally available. It is shown that in such a case direct *ab initio* solution of protein structures is feasible. Tests based on calculated diffraction data suggest the procedure to follow for a possible success.

Notation

$F_p = \|F_p\| \exp(i\varphi)$	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 (added to the native protein)
$\Phi = \varphi_{\mathbf{h}} - \varphi_{\mathbf{k}} - \varphi_{\mathbf{h}-\mathbf{k}}$	
$E_p = R \exp(i\varphi)$	Normalized structure factor for the protein
$E_d = S \exp(i\psi)$	Normalized structure factor for the isomorphous derivative
N	Number of non-H atoms in the primitive cell
$\sigma_i = \sum_{j=1}^{N} Z_j^i$	(Z_j is the atomic number of the 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 $N_{\rm eq}$ relative to the heavy-atom structure

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

† Present address: Department of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, England.

$G = 2|R_{\mathbf{h}}R_{\mathbf{k}}R_{\mathbf{h}-\mathbf{k}}|[\sigma_3/\sigma_2^{3/2}]_p$

f_j Atomic scattering factor of the jth atom

$\sum_H = \sum_H f_j^2$ (The sum is extended to the heavy-atom structure)

$D_i(x) = I_i(x)/I_o(x)$ (I_i is the modified Bessel function of order i)

Introduction

Are traditional direct methods able to solve protein structures *ab initio*? Reasons for failure are today well documented: (*a*) the weak correlation between the reliability parameter G and the value of Φ [*i.e.* flat distributions $P(\Phi|G)$]; (*b*) the low resolution of the experimental data, which hardly extend at atomic resolution; (*c*) the enormous number of local maxima for the tangent formula, which are bereft of structural meaning. However some *a posteriori* trials (Woolfson & Yao, 1990; Sheldrick, Danter, Wilson, Hope & Sieker, 1993) on previously solved small proteins succeeded in two cases (the 0.98 Å data for APP, a 36-residue hormone crystallizing in $C2$, and rubredoxin from *Desulfovibrio vulgaris*, also diffracting at atomic resolution) and excited new interest in future developments.

The question of the successful application of traditional direct methods to proteins may be answered provided two basic problems are solved. (1) Can some criteria be fixed for predicting or excluding *a priori* the success of direct methods when applied to a given set of diffraction data? (2) Under which conditions can the 'correct solutions' be picked up among numerous trials in a multisolution approach? An answer to both these questions has recently been given by Giacovazzo, Guagliardi, Ravelli & Siliqi (1994). Their main conclusions are:

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

$$z_{\mathbf{h}} = \langle \alpha_{\mathbf{h}} \rangle / \sigma_{\alpha_{\mathbf{h}}} \qquad (1)$$

may be considered to be a 'signal-to-noise ratio'. $\alpha_{\mathbf{h}}$ is the well known reliability parameter connected with the tangent formula (Karle & Hauptman, 1956).

$$\tan\theta_{\mathbf{h}} = \sum_{j=1}^{r} G_j \sin(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j}) \Big/ \sum_{j=1}^{r} G_j \cos(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j}) = T_{\mathbf{h}}/B_{\mathbf{h}}. \qquad (2)$$

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

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

(*b*) Since $\alpha_{\mathbf{h}}$ is normally distributed (Cascarano, Giacovazzo, Burla, Nunzi & Polidori, 1984) about

$$\langle \alpha_{\mathbf{h}} \rangle = \sum_{j=1}^{r} G_j D_1(G_j), \qquad (4)$$

with variance given by

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

traditional direct methods can successfully be applied to a given set of data if, for a sufficiently high percentage of large normalized structure factors,

$$z \geq T,$$

T is a threshold that, as a rule of thumb, can be reasonably fixed to about 3.

Criterion (*b*) (from now on referred to as the statistical solvability criterion) can easily be applied to proteins, where the G_j's are very small. In this case,

$$D_1(G_j) \simeq G_j/2, \quad \langle \alpha_{\mathbf{h}} \rangle = \sum_{j=1}^{r} G_j^2/2 \simeq \sigma^2_{\alpha_{\mathbf{h}}}$$

and $z_{\mathbf{h}} \simeq \langle \alpha_{\mathbf{h}} \rangle^{1/2}$.

Roughly speaking, the solvability criterion requires $\langle \alpha_{\mathbf{h}} \rangle$ to be larger than 9 for a large percentage of strong reflections. This situation may only occur for high-resolution data and/or for small proteins.

We show in Fig. 1 the distribution of the z values [*i.e.* the $P(z)$ curves] calculated from the experimental data for the proteins quoted in Table 1. For useful comparison, the $P(z)$ distribution of a small-molecule structure (WINTER) is also drawn. Details of the protocol used for calculating the curves in Fig. 1 are given in Table 2.

Fig. 1 confirms the pessimistic conclusions on the role of traditional direct methods drawn by Giacovazzo, Guagliardi, Ravelli & Siliqi (1994). Both experimental limits (*i.e.* the resolution of experimental data) and structure complexity make most of the proteins absolutely unsolvable *ab initio* by traditional direct methods. Only very small

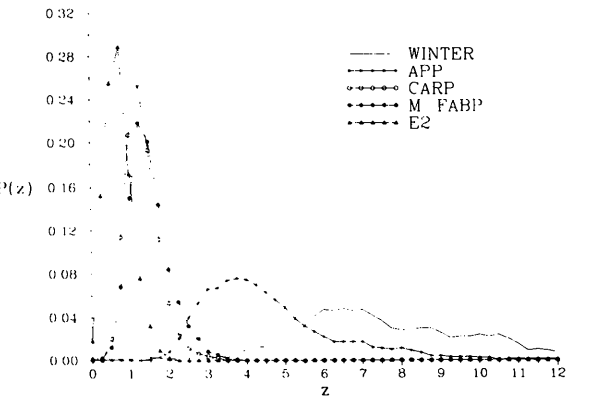

Fig. 1. The distribution of the z values calculated from experimental data for the test structures. The Cochran parameter is used in the z expression.

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

Structure code	Reference	Space group	Molecular formula	Z
APP	(1)	$C2$	$C_{190}N_{53}O_{58}Zn$	4
CARP	(2)	$C2$	$C_{513}N_{131}Ca_2O_{121}S$	4
E2	(3)	$F432$	$C_{1170}N_{310}O_{366}S_7$	96
M-FABP	(4)	$P2_12_12_1$	$C_{667}N_{170}O_{261}S_3$	4
WINTER	(5)	$P2_1$	$C_{52}H_{83}N_{11}O_{16}.3CH_2Cl_2$	2

References: (1) Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983); (2) Kretsinger & Nockolds (1973); (3) Mattevi, Obmolova, Schulze, Kalk, Westphal, De Kok & Hol (1992); (4) Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992); (5) Butters, Hütter, Jung, Pauls, Schmitt, Sheldrick & Winter (1981).

Table 2. *Protocol used for calculating the curves in Fig. 2*

RES [$=\lambda/(2\sin\theta_{max})$] is the resolution of diffraction data for the native protein, NREFL is the number of measured symmetry-independent reflections, NLAR is the number of largest normalized structure factors, NTRIP is the number of triplets found among the NLAR reflections. For the test proteins, NLAR is chosen so as to give rise to approximately 30000 triplet relationships. It is supposed the NLAR reflections are uniformly distributed in the resolution ranges: no care is taken about their centric or non-centric nature.

Structure code	RES (Å)	NREFL	NLAR	NTRIP
APP	0.99	17058	1250	30000
CARP	1.70	5056	800	28260
E2	3.00	10388	600	30000
M-FABP	2.14	7804	800	30000
WINTER	0.84	6509	475	5948

proteins (like APP) could in favourable conditions be directly phased. A countercheck for these conclusions is Table 3, where statistical calculations on triplet invariants are made. In the table, Nr is the number of triplets having G larger than ARG, % is the percentage of the positive cosine triplets and $\langle|\Phi|\rangle$ is the average of the absolute values of Φ. While triplets for APP show a favourable (for a possible successful direct phasing) behaviour, their distributions for CARP, E2 and M-FABP are close to random.

The statistical solvability criterion is of high relevance when first principles must be fixed. For example, the dogmatic principle 'direct methods do not work when atomic resolution is not attained' is not supported by our criterion, which solves the problem on a practical basis; when only low-resolution data are available, the signal-to-noise ratio is too small for proteins. However, small-molecule structures can in principle be solved even at non-atomic resolution and, in addition, proteins could in principle be solved *ab initio via* low-resolution data provided some supplementary information allowing more accurate estimates (*i.e.* higher $|G|$ values) for triplet invariants is available.

Table 3. *Statistical calculations for triplet invariants (native proteins) estimated by the Cochran formula*

APP				CARP			
ARG	Nr	%	$\langle\|\Phi\|\rangle$	ARG	Nr	%	$\langle\|\Phi\|\rangle$
0.0	30000	70.7	65.811	0.0	28260	56.8	82.352
0.2	30000	70.7	65.811	0.2	11497	58.9	80.436
0.4	30000	70.7	65.811	0.4	269	65.4	73.178
0.8	10275	73.2	62.705	0.8	0		
1.2	1025	77.7	57.678	1.2	0		
2.0	18	83.3	53.056	2.0	0		
E2				M-FABP			
ARG	Nr	%	$\langle\|\Phi\|\rangle$	ARG	Nr	%	$\langle\|\Phi\|\rangle$
0.0	30000	52.4	87.502	0.0	30000	54.6	84.764
0.2	233	53.6	85.021	0.2	15751	55.5	83.616
0.4	0			0.4	569	56.9	80.657
0.8	0			0.8	0		
1.2	0			1.2	0		
2.0	0			2.0	0		

A new question now arises: can direct methods solve protein structures if some additional prior information is available? Several examples can be found in the literature where direct methods are successfully used for phase expansion (*i.e.* from a subset of *a priori* determined phases to a larger set of phases) or for phase refinement. Since we are interested in the *ab initio* solution,, typical additional information to consider may be that contained in one or more isomorphous data sets or in measurements of the anomalous-dispersion effect. We here focus our attention on the first case. Accordingly, the question may be restated: are direct methods able to solve protein structures *ab initio* when diffraction data from an isomorphous derivative are available? Can some criteria be fixed that predict or exclude success in these new conditions?

If, besides protein intensity data, one set of isomorphous data is also available, a mathematical technique can be used (Hauptman, 1982) that integrates direct-methods and isomorphous-replacement techniques. The triplet phase invariants of the protein may then be estimated *via* the following probabilistic formula:

$$P(\Phi|R_1,R_2,R_3,S_1,S_2,S_3)\simeq[2\pi I_o(A)]^{-1}\exp(A\cos\Phi), \tag{5}$$

where A is a positive or negative term, the value of which depends on an intricate interrelationship among the six moduli R_1, R_2, R_3, S_1, S_2 and S_3. Hauptman's approach has been reconsidered and generalized by Giacovazzo, Cascarano & Zheng (1988). When the isomorphous derivative is obtained by addition of some heavy atoms, a simplified expression for A comes out:

$$A=2[\sigma_3/\sigma_2^{3/2}]_p R_1R_2R_3+(\Sigma_{3H})(|F_{d_1}|-|F_{p_1}|)\times(|F_{d_2}|-|F_{p_2}|)(|F_{d_3}|-|F_{p_3}|), \tag{6}$$

where $2[\sigma_3/\sigma_2^{3/2}]_p R_1 R_2 R_3$ is the classical Cochran (1955) concentration parameter relative to the protein structure and

$$\Sigma_{3H} = \left[\sum_H f_j(\mathbf{h}) f_j(\mathbf{k}) f_j(\mathbf{h}-\mathbf{k})\right] \times \left\{\left[\sum_H f_j^2(\mathbf{h})\right]\left[\sum_H f_j^2(\mathbf{k})\right]\left[\sum_H f_j^2(\mathbf{h}-\mathbf{k})\right]\right\}^{-1}.$$

The summations in Σ_{3H} are extended over the heavy atoms to the native protein. In terms of normalized and pseudonormalized structure factors, (6) may be written as

$$A = 2[\sigma_3/\sigma_2^{3/2}]_p R_1 R_2 R_3 + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_1 \Delta_2 \Delta_3, \quad (7)$$

where $\Delta = (|F_d| - |F_p|)/(\Sigma_H)^{1/2}$ is a pseudonormalized difference (with respect to the heavy-atom structure). Since $[\sigma_3/\sigma_2^{3/2}]_H >> [\sigma_3/\sigma_2^{3/2}]_p$, the Cochran parameter is often negligible with respect to the term including the pseudonormalized differences, and this last may attain large values even for large protein structures. Since the product $\Delta_1\Delta_2\Delta_3$ may be positive or negative, positive as well as negative triplets can be identified *via* (5). In the phasing process, a modified tangent formula can then be applied, according to which the most probable value of $\varphi_\mathbf{h}$ is given by

$$\tan\theta_\mathbf{h} = \sum_{j=1}^{r} A_j \sin(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j}) \Big/ \sum_{j=1}^{r} A_j \cos(\varphi_{\mathbf{k}_j} + \varphi_{\mathbf{h}-\mathbf{k}_j}) = T'_\mathbf{h}/B'_\mathbf{h}. \quad (8)$$

The reliability parameter is now

$$\alpha_\mathbf{h} = (T'^2_\mathbf{h} + B'^2_\mathbf{h})^{1/2}, \quad (9)$$

which is expected to be larger than the value provided by (3). Accordingly,

$$\langle\alpha_\mathbf{h}\rangle = \sum_{j=1}^{r} |A_j D_1(A_j)|$$

and

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

In these conditions, the parameter $z_\mathbf{h} = \langle\alpha_\mathbf{h}\rangle/\sigma_{\alpha_\mathbf{h}}$ may again be considered as a signal-to noise ratio. If the distribution $P(z)$ satisfies the criterion (*b*), that would suggest a possible success for a direct phasing procedure. We check this point in the next section of this paper. First, we show that a sounder parameter A can be found.

A general probabilistic formula

The concentration parameter A of the distribution (5) was derived by Giacovazzo, Cascarano & Zheng (1988) (see that paper for the notation) as

$$\begin{aligned} A = {} & 2\beta_0 R_1 R_2 R_3 + 2\beta_{11} S_1 R_2 R_3 T_1 \\ & + 2\beta_{12} R_1 S_2 R_3 T_2 + 2\beta_{13} R_1 R_2 S_3 T_3 \\ & + 2\beta_{23} S_1 S_2 R_3 T_1 T_2 + 2\beta_{22} S_1 R_2 S_3 T_1 T_3 \\ & + 2\beta_{21} R_1 S_2 S_3 T_2 T_3 + 2\beta_3 S_1 S_2 S_3 T_1 T_2 T_3, \end{aligned} \quad (10)$$

where

$$T_i = D_1(2\beta_{0i} R_i S_i).$$

Expressions (6) and (7) were obtained from (10) for the case of a native-protein heavy-atom derivative and on the assumption that

$$T_i = 1 \quad \text{for } i = 1,2,3.$$

This last assumption is strictly valid if the scattering power of the heavy atoms added to the native protein is negligible compared with the protein scattering power. In the most general case, (6) and (7) should be replaced by

$$\begin{aligned} A &= 2[\sigma_3/\sigma_2^{3/2}]_p R_1 R_2 R_3 + 2(\Sigma_{3H})(|F_{d_1}|T_1 - |F_{p_1}|) \times (|F_{d_2}|T_2 - |F_{p_2}|)(|F_{d_3}|T_3 - |F_{p_3}|) \\ &\simeq 2[\sigma_3/\sigma_2^{3/2}]_p R_1 R_2 R_3 + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta'_1 \Delta'_2 \Delta'_3, \end{aligned} \quad (11)$$

where

$$\Delta' = (|F_d|T - |F_p|)/\Sigma_H^{1/2}.$$

In (11), $|F_d|$ is multiplied by T before the calculation of the pseudonormalized difference. Accordingly, Δ' and Δ may have opposite sign and thus their use can give rise to different estimates. Even if the use of Δ' is theoretically more advisable than the use of Δ, for the cases of practical interest (*i.e.* for typical protein derivatives and for R and S larger than or close to unity), T_1 is sufficiently close to 1. Therefore, no remarkable differences in the accuracy of the results have been found whether T is used or not.

Table 4. *Parameters defining protocol for calculations*

RES [$= \lambda/(2\sin\theta_{max})$] is the resolution of the derivative diffraction data, Deriv. denotes the atomic species added to the protein, NREFL is the number of measured symmetry-independent reflections, NLAR is the number of largest normalized structure factors and NTRIP is the number of triplets found among the NLAR reflections. NLAR is chosen so as to give rise to approximately 30000 triplet relationships. $[\sigma_2]_H/[\sigma_2]_p$ is the ratio between the scattering power of the heavy atoms added to the protein and the scattering power of the protein.

Structure code	Deriv.	$[\sigma_2]_H/[\sigma_2]_p$	RES	NREFL	NLAR	NTRIP
APP	Hg	0.4555	2.0	2086	600	31807
CARP	Hg	0.0877	2.0	4687	800	30026
E2	Hg	0.0770	3.0	9179	450	35702
M-FABP	Hg	0.0642	3.0	3069	600	33110

Preliminary tests

The robustness of a phasing method has always to be checked with experimental data. Indeed, a mathematical theory, even if correct, fails if it exacts an accuracy level for the experimental data that is not attainable in practice. This is the key to the success of traditional direct methods when applied to small molecules. When the classical tangent formula (2) is used, the reliability parameter G depends on the product of three R magnitudes alone: in this case, even experimental errors up to 15–20% in R would not change the general effectiveness of the formula. As a practical counterpart, small-molecule structures are easily solved by traditional direct methods even if remarkable errors in the diffraction measurements or in their treatment have been made.

When native and isomorphous data are simultaneously available and (5) has to be used, then the Δ' (or Δ) magnitudes must be considered together with the factor $[\sigma_3/\sigma_2]_H^{3/2}$. This last term is not known *a priori* and may only be estimated. Furthermore, the terms Δ' are very sensitive to two things: (*a*) experimental errors in the measured data (even small errors in R and S can change the sign of Δ'); (*b*) imperfect treatment of the data. For example, let us suppose that R and S are obtained *via* a Wilson plot followed by a normalization process. Errors in the absolute scale and in the thermal factors for protein and derivative data can again modify the sign and value of Δ'; (*c*) lack of isomorphism between native and

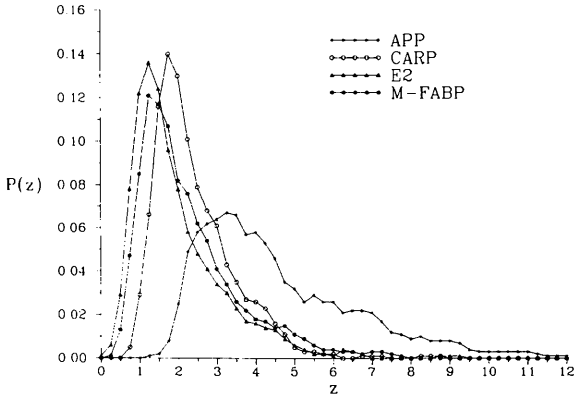

Fig. 2. The distribution of the z values (from error-free calculated data) for the test structures when A [as defined by (11)] is used in the z expression.

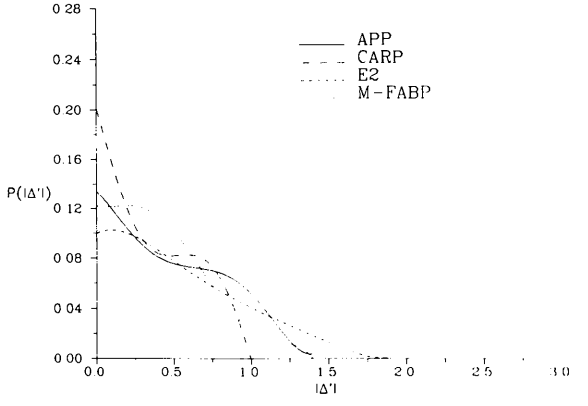

Fig. 3. Distributions of the $|\Delta|$ values (from error-free calculated data) for the NLAR reflections defined in Table 4.

Table 5. *Statistical calculations for triplet invariants estimated via* (11)

Calculated data for native and derivative structures are used. Nr is the number of triplets having $|A| > |\mathrm{ARG}|$, % is the percentage of triplets whose cosine sign is correctly estimated and $\langle|\Phi|\rangle$ is the average of the absolute values of the triplet phase Φ.

APP

	Positive estimated triplets			Negative estimated triplets		
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	29183	62.5	75.731	2624	77.4	120.760
0.2	23937	64.9	73.029	1556	83.7	127.853
0.4	10745	73.0	63.307	941	88.4	133.233
0.8	2718	87.3	46.170	369	95.9	141.724
1.2	1113	92.9	39.438	142	99.3	150.000
1.6	491	96.5	34.756	53	100.0	155.245
2.0	193	97.9	29.617	14	100.0	161.357
2.6	57	98.2	18.860	1	100.0	180.000
3.2	15	100.0	7.933	0		
3.8	3	100.0	.000	0		

CARP

	Positive estimated triplets			Negative estimated triplets		
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	28105	61.2	77.191	1921	77.7	122.285
0.2	16892	65.6	72.180	860	87.9	132.295
0.4	4687	79.5	56.483	430	94.4	139.981
0.8	901	91.5	42.121	94	100.0	151.500
1.2	214	94.4	40.131	12	100.0	168.833
1.6	44	88.6	38.250	1	100.0	180.000
2.0	11	81.8	41.545	0		
2.6	2	50.0	93.000	0		

E2

	Positive estimated triplets			Negative estimated triplets		
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	31207	53.9	85.310	4495	56.0	96.836
0.2	4795	62.5	76.124	1186	61.5	103.896
0.4	1269	70.1	67.381	466	67.2	110.354
0.8	256	82.0	50.762	112	72.3	118.884
1.2	62	87.1	43.887	40	77.5	122.675
1.6	20	95.0	35.600	13	84.6	126.615
2.0	12	91.7	36.000	4	100.0	139.500
2.6	3	100.0	35.000	1	100.0	170.000
3.2	1	100.0	34.000	1	100.0	170.000

M-FABP

	Positive estimated triplets			Negative estimated triplets		
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	28577	57.1	81.839	4533	70.3	133.887
0.2	8924	67.7	69.133	1407	86.2	131.866
0.4	2485	84.0	50.311	644	92.1	138.258
0.8	530	97.4	34.358	177	98.9	146.486
1.2	176	99.4	29.568	50	100.0	152.920
1.6	46	100.0	21.652	14	100.0	165.071
2.0	20	100.0	16.250	7	100.0	171.429
2.6	3	100.0	9.667	0		
3.2	1	100.0	.000	0		

derivative structures owing to rotation or translation of some structural regions.

Further difficulties for a successful phasing process are generated by the fact that resolution of the derivative data is lower than that of the native protein. This can limit the number of triplets reliably estimated by (5). Since the term A may simultaneously be affected by different sources of errors, we prefer in a first step to check the goodness of the approach in ideal conditions, that is by using calculated (error-free) data. The analysis of the results will allow us to identify the most critical points of the method and derive useful suggestions for a second paper, where experimental data will be used and the full phasing procedure will be described. Here, calculated data up to the experimental derivative resolution are used.

The protocol for the calculations is defined by the parameters shown in Table 4. The statistical solvability criterion should hold also in the case in which isomorphous data are additionally available. Therefore, in order to judge the possible success of (8) and (9), we calculate again the $P(z)$ curves (see Fig. 2). A comparison with Fig. 1 suggests that:

(*a*) the z value of a relatively high percentage of reflections for CARP, E2 and M-FABP is below 2 in both Fig. 1 and Fig. 2;

(*b*) the only remarkable improvement generated by the additional use of derivative data is the longer right tail of the curves in Fig. 2.

Our conclusion is that the additional use of derivative data improves the z values for only a limited percentage of the NLAR reflections, while the majority of them obtain marginal benefit. This statement is supported by Table 5 where a statistical check on the triplet reliability is made. Comparison of Table 5 with Table 3 shows that triplet reliability is markedly improved when (5) is used: reliably estimated positive and negative triplets are now available even for CARP, E2 and M-FABP. The statistical behaviour of $\langle|\Phi|\rangle$ is similar to that shown by small-molecule structures but for an important detail: too large a percentage of triplets with small $|A|$ values. For example, for APP $|A| < 0.2$ for 6314 triplets; the corresponding values for CARP, E2 and M-FABP are 12274, 29721 and 22779, respectively. This anomalous triplet distribution is responsible for the relatively high percentage of reflections with $z < 2$ for

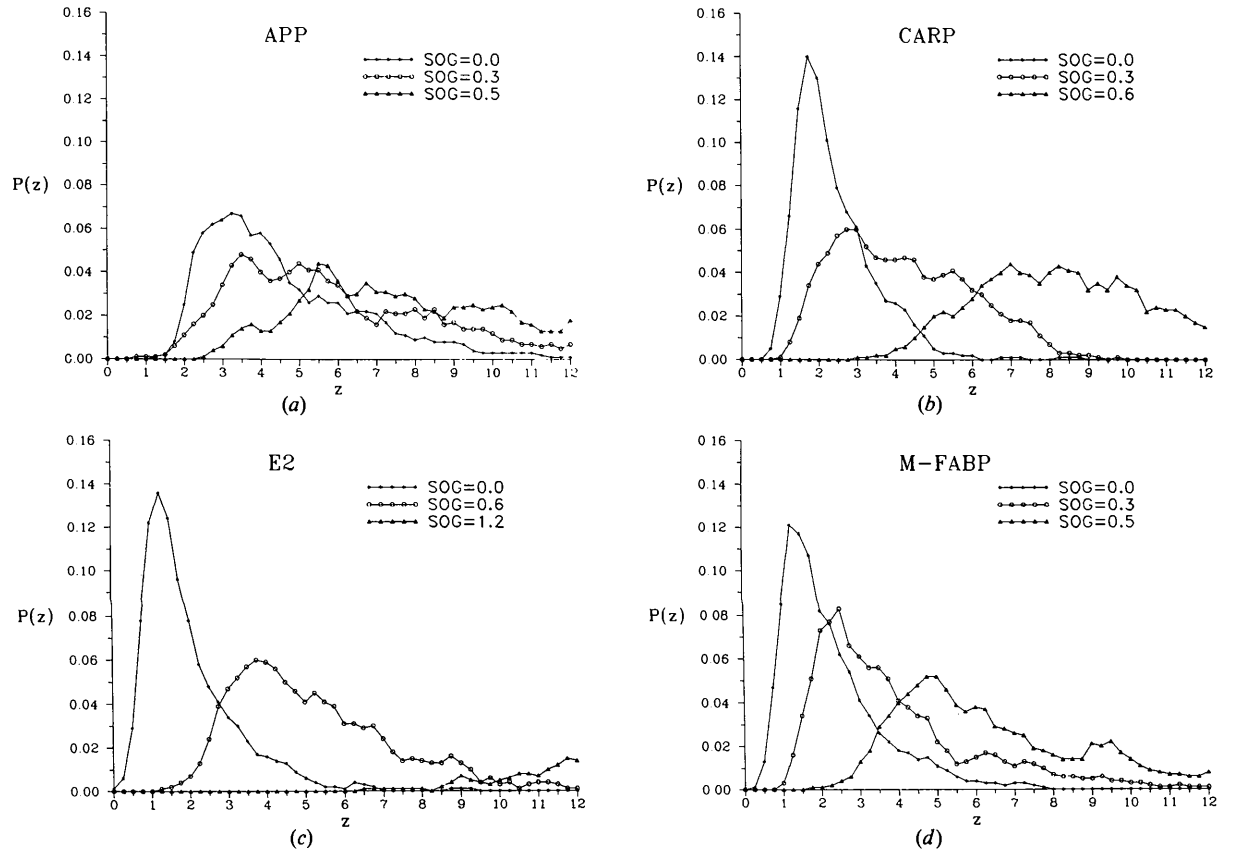

Fig. 4. Distributions of the z values (from error-free calculated data) relative to suitable sets of NLAR reflections defined by various SOG values: (*a*) APP; (*b*) CARP; (*c*) E2; (*d*) M-FABP.

Table 6. *Statistical calculations for triplet invariants estimated via* (11) *for various values of* SOG $\neq 0$

Calculated data for native and derivative structures are used.

	Positive estimated triplets			Negative estimated triplets		
APP (SOG = 0.3)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	19703	73.6	62.749	11404	73.4	116.667
0.2	17179	76.8	58.885	6984	81.0	125.219
0.4	13872	80.9	54.145	3977	86.3	131.339
0.8	5823	89.4	43.805	1362	94.5	140.361
1.2	2313	94.3	37.980	445	98.7	149.889
1.6	906	97.8	31.681	155	100.0	157.219
2.0	341	99.1	25.660	42	100.0	166.048
2.6	84	100.0	16.357	2	100.0	179.500
3.2	21	100.0	8.952	0		
3.8	3	100.0	.000	0		
APP (SOG = 0.5)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	19006	85.6	48.918	15463	82.5	127.297
0.2	18920	85.7	48.724	14571	83.8	128.868
0.4	18391	86.2	48.124	11127	86.5	132.378
0.8	10887	90.4	43.222	4196	92.8	139.692
1.2	4311	94.9	37.575	1236	97.9	147.482
1.6	1487	98.2	31.319	329	100.0	154.489
2.0	459	99.3	25.765	74	100.0	164.946
2.6	96	100.0	16.323	3	100.0	179.667
3.2	21	100.0	8.952	0		
3.8	3	100.0	.000	0		
CARP (SOG30.3)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	16919	81.1	54.649	12281	78.4	122.973
0.2	15825	83.3	52.313	8002	85.2	130.771
0.4	12405	87.5	47.149	4225	91.2	136.871
0.8	4056	95.9	37.613	1007	99.7	148.695
1.2	867	99.5	30.278	139	100.0	159.338
1.6	103	100.0	19.417	6	100.0	173.000
2.0	7	100.0	11.000	0		
2.6	0			0		
CARP (SOG = 0.6)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	20419	96.4	35.004	17574	95.9	144.161
0.2	20419	96.4	35.004	17564	95.9	144.166
0.4	20418	96.4	34.999	17471	95.9	144.233
0.8	17078	97.5	33.119	10023	98.8	148.645
1.2	4751	100.0	27.265	1718	100.0	157.194
1.6	491	100.0	17.094	130	100.0	166.246
2.0	7	100.0	10.286	0		
2.6	0			0		

	Positive estimated triplets			Negative estimated triplets		
E2 (SOG = 0.6)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	16027	64.8	73.048	15531	62.0	104.210
0.2	15536	65.1	72.605	11293	65.1	107.659
0.4	9082	69.5	67.370	5268	71.1	115.061
0.8	2027	79.8	54.358	1174	82.5	126.718
1.2	506	87.4	45.103	339	89.4	135.973
1.6	158	90.5	41.101	114	94.7	142.316
2.0	62	90.3	37.161	35	91.4	140.143
2.6	12	100.0	28.750	11	90.9	155.000
3.2	4	100.0	16.500	6	100.0	172.883
3.8	2	100.0	12.000	2	100.0	180.000
4.4	1	100.0	1.000	0		
E2 (SOG = 1.2)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	17647	86.2	47.580	17123	85.2	131.182
0.2	17647	86.2	47.580	17123	85.2	131.182
0.4	17647	86.2	47.580	17123	85.2	131.182
0.8	17595	86.2	47.542	16880	85.4	131.371
1.2	12084	88.9	44.261	10671	88.8	135.639
1.6	5396	91.7	39.951	4654	92.2	140.573
2.0	2143	93.0	36.874	1849	93.6	143.607
2.6	617	95.0	32.968	567	97.7	150.026
3.2	209	98.1	26.646	198	96.0	151.172
3.8	89	96.6	25.360	80	96.2	152.137
4.4	28	96.4	16.607	33	100.0	160.061
M-FABP (SOG = 0.3)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	17408	72.9	63.667	13987	72.1	115.612
0.2	14669	76.6	59.318	6414	85.5	130.980
0.4	7038	87.4	46.786	2891	91.7	138.582
0.8	1651	96.4	34.681	806	96.9	145.949
1.2	526	99.2	29.717	258	99.6	152.519
1.6	169	99.4	23.959	105	100.0	157.486
2.0	67	100.0	19.269	44	100.0	164.909
2.6	12	100.0	11.750	9	100.0	170.000
3.2	1	100.0	.000	3	100.0	171.667
M-FABP (SOG = 0.5)						
|ARG|	Nr	%	$\langle\|\Phi\|\rangle$	Nr	%	$\langle\|\Phi\|\rangle$
0.0	16854	88.6	45.423	15978	87.0	133.076
0.2	16852	88.6	45.419	15554	87.6	133.730
0.4	15091	90.5	43.199	11635	91.7	138.618
0.8	6006	95.9	35.499	4056	96.5	145.359
1.2	2050	98.7	30.927	1332	99.2	150.208
1.6	707	99.6	26.320	523	100.0	155.589
2.0	258	100.0	22.705	188	100.0	159.048
2.6	58	100.0	18.034	36	100.0	163.889
3.2	10	100.0	6.300	9	100.0	174.333

CARP, E2 and M-FABP, and therefore for a probably difficult phase expansion in a direct-phasing process. The primary source of this undesired effect is an intrinsic property of the distribution of the $|\Delta'|$ values. In Fig. 3, for each structure, the experimental distributions of the $|\Delta'|$ values for the NLAR reflections defined in Table 4 are given. The curves suggest that the most probable value of $|\Delta'|$ is close to zero: therefore, too small $|\Delta'|$ values should be associated with a relatively high percentage of the NLAR reflections. When $|\Delta_{\mathbf{h}}|$ is small, the reflection $\mathbf{h}$ is likely to be characterized by a small value of $\alpha_{\mathbf{h}}$; consequently, $z_{\mathbf{h}}$ will also be small and the estimate of $\varphi_{\mathbf{h}}$ will be unreliable. We reacted to this unfavourable situation by changing the nature of the NLAR reflections (but leaving unmodified the value of NLAR): we included in the set the reflections with the largest R values provided $|\Delta'| >$ SOG, where SOG is a suitable threshold. The condition $|\Delta'| >$ SOG selects the reflections whose phase values may be reliably estimated (in a probabilistic sense); the condition 'R large' is dictated by the opportunity of obtaining a valuable contribution to the Fourier synthesis once the reflection is phased. The distributions $P(z)$ are now recalculated for the new set of NLAR reflections. Curves corresponding to various

values of SOG are shown in Figs. 4(*a*)–(*d*) for each test structure.

Curves corresponding to SOG = 0 coincide with those displayed in Fig. 2 and are quoted again in Fig. 4 for the benefit of the reader. It is easily seen that: (*a*) the curves shift remarkably to the right when SOG increases; (*b*) the percentage of reflections with $z < 3$ progressively decreases for higher values of SOG and soon becomes negligible. Accordingly, the curves show very long right tails, suggesting a high percentage of reliably estimable phases. In order to check the above conclusions, we show in Table 6 the overall triplet statistics for the various test structures and for the SOG ≠ 0 values used in Fig. 4.

The comparison of Table 6 with Table 5 immediately suggests two things.

(*a*) The overall reliability of the estimates increases with SOG.

(*b*) The number of unreliable triplets progressively comes down for higher SOG values. For example, for E2, 29721 triplets have $|A| < 0.2$ when SOG = 0.0; this number reduces to 4729 when SOG = 0.6 and to zero when SOG = 1.2. A similar trend is found for all the other test structures.

The above results suggest that almost all the reflections in the set NLAR could reliably be estimated by a direct-phasing process. However, one issue remains open: the minimum value of R among the NLAR reflections (say R_{min}) decreases when SOG increases. R_{min} could then be so small that several reflections, once phased, would negligibly contribute to Fourier syntheses. In Table 7, the value of R_{min} is shown for each test structure and for each value of SOG. It is seen that R_{min} is sufficiently large to guarantee a useful contribution to Fourier syntheses for each of the NLAR reflections.

Table 7. *The value of R_{min} found for the various test structures among the* NLAR *reflections when some values of* SOG *are used*

Structure code	NLAR	SOG	R_{min}
APP	600	0.0	1.09
		0.3	0.86
		0.5	0.60
CARP	800	0.0	1.29
		0.3	1.07
		0.6	0.87
E2	450	0.0	1.81
		0.6	1.45
		1.2	0.85
M-FABP	600	0.0	1.24
		0.3	1.04
		0.5	0.73

Concluding remarks

We have examined the question: 'is a protein structure solvable *ab initio* by direct methods when diffraction data from one isomorphous derivative are additionally available?' The application of the statistical solvability criterion to calculated (error-free) diffraction data suggests a positive answer, provided the set of reflections to be actively used in the phasing process is characterized by relatively high values of $|E|$ *and* $|\Delta|$. Complementary tests on the overall reliability of the triplet invariant estimates confirm what is suggested by the solvability criterion.

The role of reflections with high $|\Delta'|$ value was already perceived by Fortier, Weeks & Hauptman (1984) and by Karle (1983). This paper shows how crucial they are for the success of the phasing process and provides experimental details about their use. As a consequence, direct procedures designed for small molecules must be greatly modified in order for one to profit from the enormous amount of information contained in the experimental data.

The application of the method to real data is now mandatory. We anticipate here that a phasing procedure has been devised that, applied to real experimental data, will allow the *ab initio* solution of all the four test protein structures used in this paper. This also confirms that high-resolution data, perfect isomorphism and negligible errors in measurements, even if desirable, are not so critical as generally believed. The phasing procedure and the experimental results will be described in paper II of this series.

References

BUTTERS, T., HÜTTER, P., JUNG, G., PAULS, N., SCHMITT, H., SHELDRICK, G. M. & WINTER, W. (1981). *Angew. Chem.* **93**, 904–905.

CASCARANO, G., GIACOVAZZO, C., BURLA, M. C., NUNZI, A. & POLIDORI, G. (1984). *Acta Cryst.* A**40**, 389–394.

COCHRAN, W. (1955). *Acta Cryst.* **8**, 473–478.

FORTIER, S., WEEKS, C. M. & HAUPTMAN, H. (1984). *Acta Cryst.* A**40**, 544–548.

GIACOVAZZO, C., CASCARANO, G. & ZHENG, C.-D. (1988). *Acta Cryst.* A**44**, 45–51.

GIACOVAZZO, C., GUAGLIARDI, A., RAVELLI, R. & SILIQI, D. (1994). *Z. Kristallogr.* **209**, 136–142.

GLOVER, I., HANEEF, I., PITTS, J., WOOD, S., MOSS, D., TICKLE, I. & BLUNDELL, T. (1983). *Biopolymers*, **22**, 293–304.

HAUPTMAN, H. (1982). *Acta Cryst.* A**38**, 289–294, 632–641.

KARLE, J. (1983). *Acta Cryst.* A**39**, 800–805.

KARLE, J. & HAUPTMAN, H. (1956). *Acta Cryst.* **9**, 635–651.

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.

SHELDRICK, G. M., DAUTER, Z., WILSON, K. S., HOPE, H. & SIEKER, L. C. (1993). *Acta Cryst.* D**49**, 18–23.

WOOLFSON, M. M. & YAO, J.-X. (1990). *Acta Cryst.* A**46**, 409–413.

ZANOTTI, G., SCAPIN, G., SPADON, P., VEERKAMP, J. H. & SACCHETTINI, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.