RESEARCH PAPERS

Acta Cryst. (1995). A**51**, 811–820

The *Ab Initio* Crystal Structure Solution of Proteins by Direct Methods. V. A New Normalizing Procedure

BY CARMELO GIACOVAZZO AND DRITAN SILIQI*

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

AND JAVIER GONZALEZ PLATAS

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

(*Received* 10 *November* 1994; *accepted* 14 *March* 1995)

Abstract

In papers I–III of this series [Giacovazzo, Siliqi & Ralph (1994). *Acta Cryst.* A**50**, 503–510; Giacovazzo, Siliqi & Spagna (1994). *Acta Cryst.* A**50**, 609–621; Giacovazzo, Siliqi & Zanotti (1995). *Acta Cryst.* A**51**, 177–188], a direct phasing method is described that proved potentially able to solve *ab initio* protein structures. The method exploits diffraction data from the native protein and one isomorphous derivative, and was tested on experimental data with resolution equal to or higher than 3 Å. More severe tests performed at 4 Å resolution on a larger set of test structures – some of which had a low-quality derivative – suggested that the normalizing procedure proposed in paper II had to be revised. A new approach is here described for normalizing data, which enables the phasing method to succeed also at low (~ 4 Å) resolution for data affected by a severe lack of isomorphism. A procedure capable of discarding trial solutions with high figures of merit but devoid of structural meaning is also described.

Symbols and abbreviations

See papers I–III (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995).

Introduction

In papers I–III of this series, a probabilistic approach has been described for the *ab initio* crystal structure solution of proteins. The method integrates direct methods and isomorphous techniques and requires diffraction data from the native protein and from one isomorphous derivative. It is based on the formula obtained by Giacovazzo, Cascarano & Zheng (1988) [see also Hauptman (1982) for a related expression] estimating three-phase invariants given six magnitudes and has been tested on experimental data of four proteins, APP, CARP, E2, M-FABP (see Table III.1). The results may be described as follows:

(1) The multisolution technique is applied to random starting planes. A small number of trials is sufficient for obtaining the correct solution.

(2) Proper figures of merit rank the trials: the correct solution is often found among the trials characterized by the largest values of the combined figure of merit CFOM.

(3) About 40% of the reflections up to derivative resolution can be phased with good reliability. The process is fast and requires about 10–20 s of CPU time on an IBM RISK 6000 mod.320E machine.

(4) The accuracy of the phasing process relies on the quality of the heavy-atom derivative.

Quite small phase errors can be obtained in the case of good isomorphism. Severe lack of isomorphism degrades the accuracy of the triplet invariant estimates and therefore the quality of the assigned phases. The process proved sufficiently robust against experimental errors but it may still be improved in several ways. In this paper, we focus our attention on the following ones:

(*a*) The procedure is based on Δ (or Δ') values, which are obtained by a statistical treatment of the experimental data. Each Δ may be considered as the sum of a signal (*i.e.* the heavy-atom scattering) and a noise (arising from the disordered water distribution, lack of isomorphism, errors in measurements *etc.*). Since noise is unavoidable in the protein, the following question arises: is the procedure yielding Δ values optimally designed in the case of large noise? If not, can new criteria be fixed to design a robust procedure accurately working in severe conditions?

(*b*) The phasing procedure was applied to cases for which derivative data up to 3 Å resolution are available (3 Å resolution for E2 and M-FABP, 2 Å resolution for

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

CARP and APP). This was an important goal: indeed, it was proved that success for direct methods can be obtained even at non-atomic resolution. However, it is not infrequent that only isomorphous data up to 4 Å resolution are available. Can the phasing process successfully work at such a low resolution where the scaling Wilson procedure is rather inaccurate?

(*c*) The various trial solutions are ranked by proper FOM's, which are extremely efficient when perfect isomorphism occurs. For real cases, one can expect that the correct solution is among the trials with the highest values of CFOM, however, various experimental errors and lack of isomorphism can greatly reduce the discriminating power of the various FOM's (see paper II). The search of FOM's less sensitive to the various 'errors' is a topic of enormous importance for the success of direct methods applied to macromolecules. An alternative way of contributing to the solution of the problem may consist in answering the following question: do criteria exist that are able to discard, among the various trials with the highest values of CFOM, the trial solutions devoid of structural meaning?

In this paper, we describe techniques that provide efficient solutions to the problems described in (*a*), (*b*) and (*c*).

The normalization process at 4 Å resolution

In order to check our phasing process for a larger variety of structures, including some more severe tests, we enlarged our set of test structures by adding BPO, FIS, NOX and TAQ to APP, CARP, E2 and M-FABP. Isomorphism for FIS, NOX and TAQ derivatives is very poor, while it is excellent for BPO.

Data for the bromoperoxidase A2 from *Streptomyces aureofaciens* ATCC 10762 (BPO) (Hecht, Sobek, Haag, Pfeifer & Van Pee, 1994) have been collected at room temperature. The space group is cubic $P2_13$ with a lattice constant $a = 126.5$ Å. The asymmetric unit contains two molecules of 30 000 Da each. Native data were measured to 2.05 Å resolution, two derivative data sets were measured to 2.6 Å resolution. The resulting MIR map was easily interpretable and allowed a complete chain tracing for both molecules in the asymmetric unit.

The factor for inversion stimulation (FIS) crystallizes as a homodimer, with 2 × 98 amino acid residues in the asymmetric unit. The two monomers are related by a non-crystallographic dyad axis. The crystal structure was determined by multiple isomorphous replacement (Kostrewa, Granzin, Stock, Choe, Labahn & Saenger, 1992). In this paper, we will only use native data (up to 2 Å resolution) and the $[PtCl_2(C_2H_4)]_2$-derivative data (resolution up to 3.3 Å).

NOX is the code name for NADH oxidase from *Thermus thermophilus* (Hecht, Erdmann, Park, Sprinzl, Schmid & Schomburg, 1993; Hecht, Erdmann, Park, Sprinzl & Schmid, 1995). Native data were collected at 285 K to 2.3 Å resolution for a total of 9274 reflexions. Potential derivative data sets were collected up to 2.9 Å resolution. Of the five derivative data sets, only the $[PtCl_2(NH_2)]$ derivative is used in this paper.

TAQ is the code name for adenine-N^6-DNA-methyltransferase. The protein consists of 421 amino acid residues and its molecular weight is 47 856 Da. It crystallizes with two molecules per asymmetric unit, which are related by a local twofold screw axis parallel to the *c* axis (Labahn, Granzin, Schluckebier, Robinson, Jack, Schildkraut & Saenger, 1995). Native data were collected up to 2.4 Å resolution. Three derivatives were used for crystal structure determination, K_2PtCl_4, $[PtCl_2(C_2H_4)]_2$, $(C_7H_5O_3)HgCl$. In this paper, only the $[PtCl_2(C_2H_4)]_2$- and $(C_7H_5O_3)HgCl$-derivative data sets will be involved in the calculations.

In Table 1, we collect the main crystallochemical data for all the test structures and, in Table 2, we show some relevant parameters for the diffraction data we used.

We first applied the normalization procedure described in paper II to TAQ by using Hg-derivative data (experimental data up to 4 Å resolution). The procedure is a two-step method: first, the standard Wilson method is applied to native protein data truncated at derivative resolution to obtain $(K_{DW})_p$ and $(B_{DW})_p$. Then, in the second step, estimates of K_d/K_p and $B_d - B_p$ are

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

Structure code	Reference	Space group	Molecular formula	*Z*
APP	(*a*)	*C*2	$C_{190}N_{53}O_{58}Zn$	4
CARP	(*b*)	*C*2	$C_{513}N_{131}O_{121}SCa_2$	4
E2	(*c*)	*F*432	$C_{1170}N_{310}O_{366}S_7$	96
M-FABP	(*d*)	$P2_12_12_1$	$C_{667}N_{170}O_{261}S_3$	4
BPO	(*e*)	$P2_13$	$C_{2744}N_{712}O_{1073}$	12
FIS	(*f*)	$P2_12_12_1$	$C_{783}N_{224}O_{1312}S_{10}$	4
NOX	(*g*)	$P4_12_12$	$C_{1034}N_{299}O_{704}P_{1/8}S_2$	8
TAQ	(*h*)	$P2_12_12$	$C_{4390}N_{1174}O_{1240}S_8$	4

References: (*a*) Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983); (*b*) Kretsinger & Nockolds (1973); (*c*) Mattevi, Obmolova, Shulze, Kalk, Westphal, De Kok & Hol (1992); (*d*) Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992); (*e*) Hecht *et al.* (1994); (*f*) Kostrewa *et al.* (1992); (*g*) Hecht *et al.* (1993), Hecht *et al.* (1995); (*h*) Labahn *et al.* (1995).

Table 2. *Relevant parameters for diffraction data of test structures*

	Native		Derivative			
Structure code	RES (Å)	NREFL	Heavy atom	$[\sigma_2]_H/[\sigma_2]_p$	RES (Å)	NREFL
APP	0.99	17058	Hg	0.23	2.00	2086
CARP	1.71	5056	Hg	0.09	1.71	4687
E2	3.00	10388	Hg	0.08	3.00	9179
M-FABP	2.14	7595	Hg	0.06	2.15	7125
BPO	2.76	16348	Au	0.06	2.76	16348
			Pt	0.06	2.76	15291
FIS	2.00	12846	Pt	0.35	3.30	2983
NOX	3.00	4295	Pt	0.08	3.00	4295
TAQ	2.40	37268	Pt	0.09	3.20	15620
			Hg	0.04	4.00	8484

obtained by a differential Wilson plot (Blundell & Johnson, 1976) through the equation

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

The scaling and thermal parameters for the derivative were called in paper II $(K_{DW})_d$ and $(B_{DW})_d$, respectively. They are used to calculate the first estimates of the Δ' values, which are then rescaled (see paper II) by the factor

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

to make the experimental distribution of $|\Delta'|$ closer to the expected one.

The results of the procedure for TAQ were in some way surprising: strongly negative values of the thermal factors were obtained by the Wilson procedure for both the native $[(B_{DW})_p = -20.81]$ and the derivative $[(B_{DW})_d = -8.55]$ data. The Wilson plot is highly non-linear and is shown in Fig. 1.

In order to check if such a result was casual or representative of a systematic behaviour of protein data at 4 Å resolution, we cut at 4 Å the data of all the other test structures. The results are shown in Table 3: the thermal factors B are all negative for the protein data $[(B_{DW})_p \ll 0]$, differences $[(B_{DW})_d - (B_{DW})_p]$ are all positive. If the same procedure is applied to the other test data up to derivative resolution, the results fit better with expectations (see Table 3 again). Indeed, positive B_p values are now obtained and the differences $[(B_{DW})_d - (B_{DW})_p]$ are again all positive (and highly correlated with the corresponding differences obtained at 4 Å resolution).

Table 3. *Scale and thermal factors for the test structures obtained by the procedure described in paper* II

Structure code	4 Å resolution data $(K_{DW})_p$ $(K_{DW})_d$	4 Å resolution data $(B_{DW})_p$ $(B_{DW})_d$	Derivative resolution data $(K_{DW})_p$ $(K_{DW})_d$	Derivative resolution data $(B_{DW})_p$ $(B_{DW})_d$
APP	0.27	−18.86	0.14	9.78
	0.34	−14.50	0.19	12.57
CARP	3.87	−22.88	2.12	5.69
	3.92	−18.60	2.22	7.69
E2	27783.82	−11.89	19178.18	9.14
	28088.40	−6.33	20429.74	11.41
M-FABP	433.78	−31.41	211.37	6.11
	65.48	−27.99	33.70	6.54
BPO (Au)	0.091	−37.56	0.037	6.71
	0.093	−34.30	0.039	8.11
BPO (Pt)	0.091	−37.56	0.037	7.30
	0.101	−37.95	0.039	7.70
FIS	0.57	−10.61	0.39	10.12
	4.30	29.71	3.57	39.64
NOX	0.036	−39.52	0.0128	10.25
	0.035	−35.92	0.0130	11.86
TAQ (Pt)	4.88	−18.09	2.94	10.92
	0.031	−11.26	0.019	16.32
TAQ (Hg)	5.23	−20.81	5.23	−20.81
	0.036	−8.55	0.036	−8.55

The reason for the 'anomalous' behaviour at 4 Å resolution can be immediately understood from Figs. 2 and 3 where Wilson plots for native APP and FIS data are shown. In each figure, Wilson plots for the data up to 4 Å resolution and up to derivative resolution are shown together with the corresponding least-squares straight lines. It is easily seen that the impressive errors in the estimated K and B values at 4 Å resolution are a consequence of Debye effects. Indeed, the radial distribution of diffracted intensities of proteins always has a trough at about 6 Å and a peak at about 4.5 Å (Richardson & Richardson, 1985). The problem is now

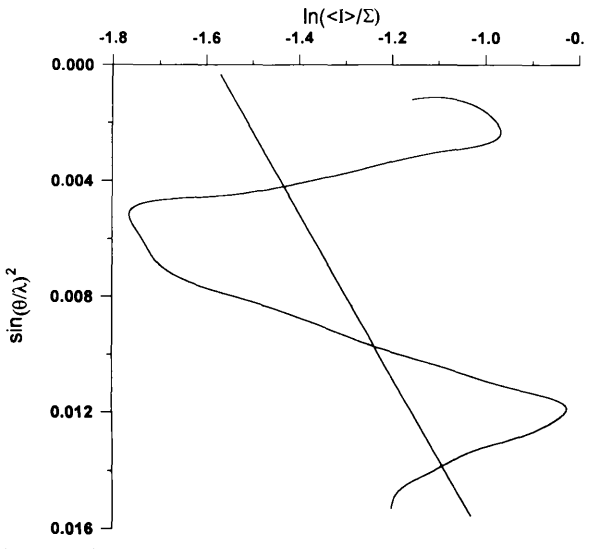

Fig. 1. TAQ: Wilson plot for native diffraction data up to derivative resolution (4 Å).

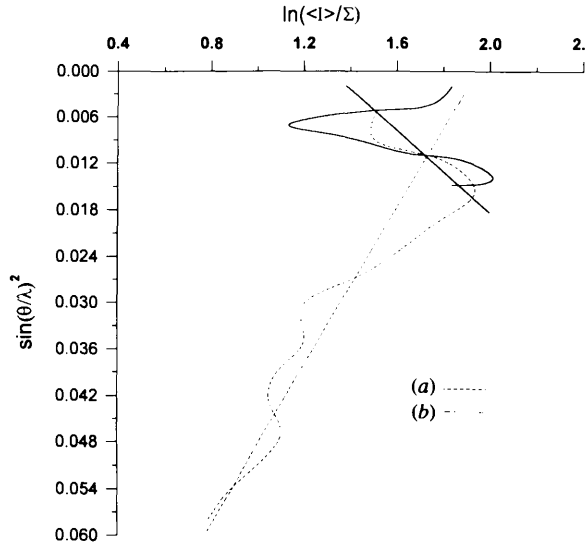

Fig. 2. APP: (*a*) Wilson plot for native diffraction data up to derivative resolution (2 Å); (*b*) Wilson plot for native diffraction data up to 4 Å resolution.

to understand if errors in the normalizing step can hinder the success of the phasing process or damage its efficiency. This really occurs for small molecules (Subramanian & Hall, 1982; Hall & Subramanian, 1982; Cascarano, Giacovazzo & Guagliardi, 1992). Does the same occur for macromolecules, where Δ (and not $|E|$) quantities are used? This is not so obvious since Δ parameters are more sensitive to the ratio K_d/K_p and to the differences $B_b - B_p$ rather than to their absolute values.

In order to answer the above question, we used Δ' values for TAQ (4 Å resolution data) to estimate, *via* (I.11), the triplet invariants among the 986 reflections with the largest values of $|\Delta'|$. Some statistics are shown in Table 4(*a*). Triplets are divided into two subsets, positive and negative estimated triplets: NR is the number of triplets having $|A| >$ 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 Φ. It is immediately seen that the number of triplets estimated negative is abnormally higher than the number of triplets estimated positive. This has no physical meaning and is mainly due to errors in the normalizing procedure. As a consequence, the percentage of correctly estimated negative triplets is smaller than 50%, and this seriously endangers the success of the phasing process. A similar result is obtained for FIS at 4 Å resolution (see Table 4*b*) using the 643 reflections with the largest values of $|\Delta'|$. Different statistics are obtained for NOX and BPO (see Tables 4*c*, *d* and *e*). For NOX, a too large percentage of the triplets found among the 564 reflections with the largest values of $|\Delta'|$ have an $|A|$ value between 0.0 and 0.2 but there is no systematic error in the estimation of the sign of the triplets. For BPO (Pt derivative), only 35 triplets have $|A| \geq 0.2$ and, for BPO (Au derivative), only 3455 have $|A| \geq 0.2$. Again, no systematic error is found in the estimation of the sign of the triplets.

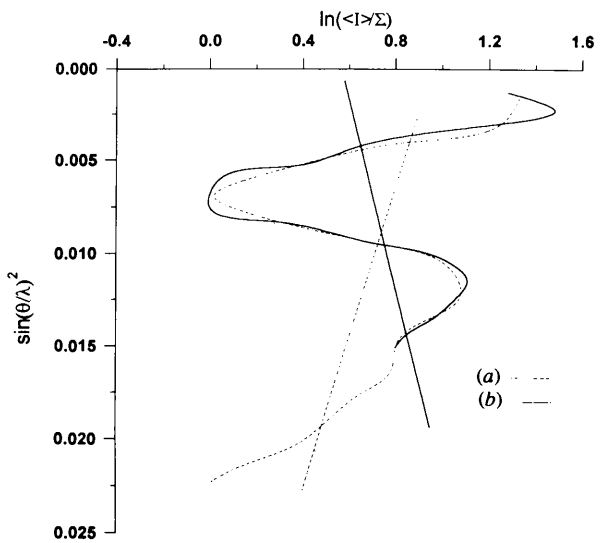

Fig. 3. FIS: (*a*) Wilson plot for native diffraction data up to derivative resolution data (3.2 Å); (*b*) Wilson plot for native diffraction data up to 4 Å resolution.

Table 4. *Statistical calculations for triplet invariants estimated via equation* (11) *of paper* I *at* 4 Å *resolution*

	Positive estimated triplets			Negative estimated triplets		
|ARG|	NR	%	$\langle\|\Phi\|°\rangle$	NR	%	$\langle\|\Phi\|°\rangle$
(*a*) TAQ (Hg derivative)						
1.6	6255	52	87	32302	48	88
2.0	5389	53	86	28693	48	88
2.6	3470	53	86	20132	48	88
3.2	1998	54	85	12620	48	88
3.8	842	57	82	5997	48	88
4.4	282	60	78	2680	47	88
5.5	70	60	82	561	47	88
(*b*) FIS						
0.4	7660	55	84	42340	49	89
0.8	5947	56	83	29277	49	89
1.2	3232	57	81	14135	48	88
1.6	1523	58	80	6144	48	88
2.0	724	56	82	2537	49	90
2.6	223	53	85	669	44	85
3.2	62	48	88	148	42	81
3.8	7	43	96	13	31	70
(*c*) NOX						
0.0	31477	54	84	18523	54	95
0.2	100	76	52	23	61	101
0.4	1	0	178			
(*d*) BPO (Au derivative)						
0.0	24151	68	69	25849	67	110
0.2	2348	78	58	1107	80	124
0.4	120	86	44	75	84	139
(*e*) BPO (Pt derivative)						
0.0	33101	64	73	16899	70	113
0.2	28	93	42	7	86	137

Table 5. *Application of the phasing procedure described in papers* II *and* III *at* 4 Å *resolution data*

Order of solution is the order of the trial solution as ranked by CFOM. NPHAS is the number of phased reflexions, ERR is the average phase error calculated with respect to published phase values.

Structure code	Order of solution	NPHAS	ERR (weighted)
APP	–	–	–
CARP	2	348	42 (37)
E2	2	1696	39 (30)
M-FABP	–	–	–
BPO (Au)	–	–	–
BPO (Pt)	–	–	–
FIS	–	–	–
NOX	–	–	–
TAQ (Pt)	1	3194	66 (57)
TAQ (Hg)	–	–	–

The above results indicate that the normalizing procedure described in paper II is unable to carefully control at 4 Å resolution the various parameters playing a role in the scaling process. This constitutes a bad premise for the success of the subsequent phase-determination procedure. This was confirmed when we applied the phasing process described in papers II and III to data up to 4 Å resolution for all the test structures. Results are

shown in Table 5. It is not a surprise that the procedure does not succeed in the majority of the cases [*i.e.* for APP, M-FABP, BPO, FIS, NOX and TAQ(Hg)]. Only E2 and CARP are satisfactorily phased, while for TAQ(Pt) a solution is found but with an appreciable mean phase error.

The question now is whether a more accurate normalizing procedure can be found that is able to overcome the difficulties met with 4 Å resolution data and possibly to improve the accuracy of the phasing process with data at higher resolution. Such a technique is described in the following sections.

The normalization procedure by histogram matching

For R and S larger than or close to unity, the factor T is so close to unity that Δ' may be replaced by Δ. The advantage of the quantity Δ is that its distribution may be readily calculated. This has been done in paper III, where, in order to guess the number of phases that should be involved in the phasing process, the probability distribution function $P(\Delta)$ has been obtained as a function of the parameter $\sigma = [\sigma_2]_H/[\sigma_2]_p$. We will see now that $P(\Delta)$ can also play a role in the normalizing process.

Let Δ_T be a positive threshold for Δ, $n^+_{\Delta_T}$ be the number of positive Δ's for which $\Delta > \Delta_T$ and $n^-_{\Delta_T}$ be the number of negative Δ for which $|\Delta| > \Delta_T$. Since $P(\Delta)$ is not an even function, the ratio

$$\mathrm{RPM} = n^+_{\Delta_T}/n^-_{\Delta_T}$$

is expected to be larger than unity for any value of σ and for any Δ_T. In Fig. 4, we show RPM curves for different values of σ. RPM increases with σ and, for a given σ, increases with Δ_T. Its value is strictly correlated with the ratio k_d/k_p: errors in the estimate of this ratio will produce anomalous values of RPM. For example, if F_d values are scaled so that they are larger than their true values, the number of positive Δ's will exceed the expected value. In the converse case, the number of negative Δ's will be larger than the expected value. In general, the experimental $P(\Delta)$ curve is modelled by different sources of error: besides the scaling error, wrong estimates of the differences $B_d - B_p$ (as a

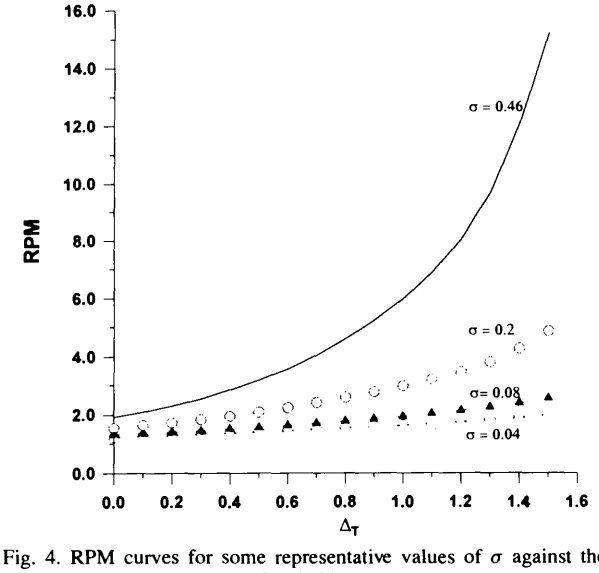

Fig. 4. RPM curves for some representative values of σ against the threshold Δ_T.

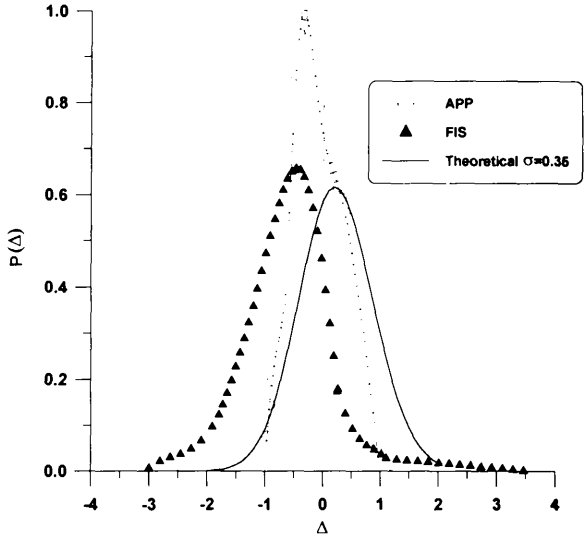

Fig. 5. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.35$ and corresponding experimental curves for APP and FIS, obtained by the normalizing procedure described in paper II (4 Å resolution).

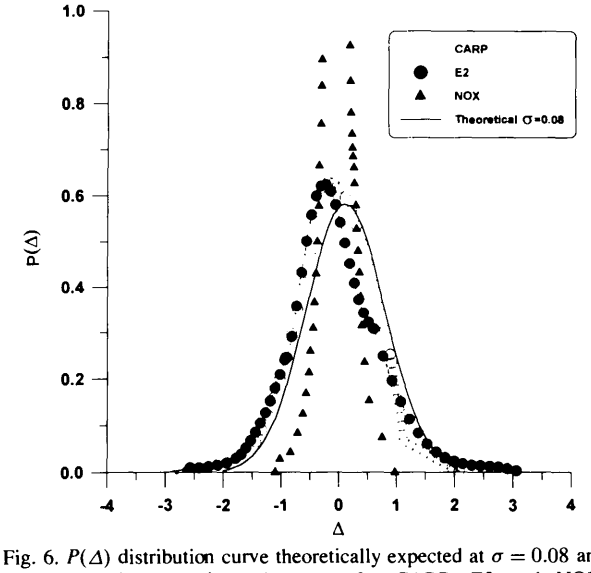

Fig. 6. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.08$ and corresponding experimental curves for CARP, E2 and NOX, obtained by the normalizing procedure described in paper II (4 Å resolution).

consequence of the scaling error), errors in measurements, lack of isomorphism *etc.* will also generate anomalies in $P(\Delta)$. It is therefore instructive to compare for all the test structures (see Figs. 5–8) the theoretical Δ curves with those obtained from mesurements at 4 Å resolution in accordance with the normalization procedure described in paper II. APP and FIS curves are shown in Fig. 5, together with the theoretical curve at $\sigma \simeq 0.35$; CARP, E2 and NOX curves are shown in Fig. 6 together with the theoretical curve at $\sigma = 0.08$; TAQ curves are shown in Fig. 7 together with the expected curve at $\sigma = 0.06$; M-FABP and BPO curves are shown in Fig. 8 together with the expected curve at $\sigma = 0.06$.

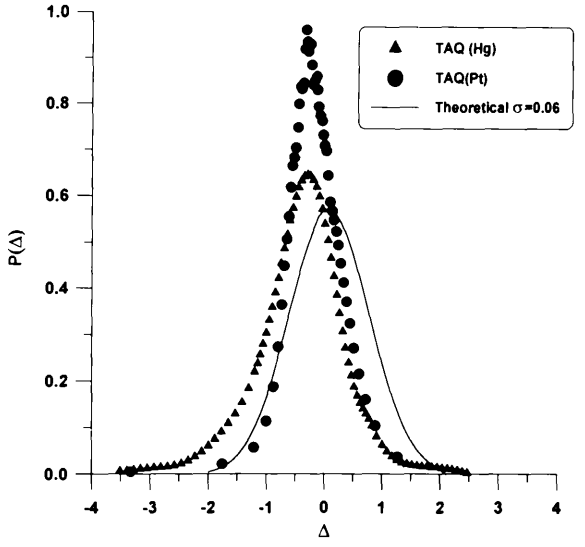

Fig. 7. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.06$ and corresponding experimental curves of TAQ (Pt derivative) and TAQ (Hg derivative), obtained by the normalizing procedure described in paper II (4 Å resolution).

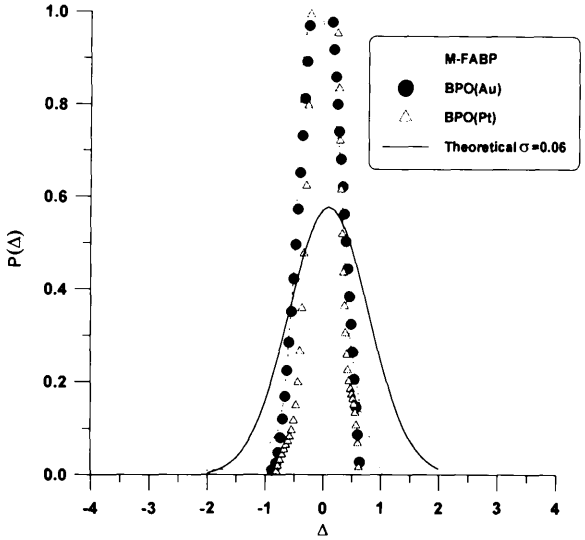

Fig. 8. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.06$ and corresponding experimental curves for BPO (Au and Pt derivatives) and M-FABP obtained by the normalizing procedure described in paper II (derivative resolution data).

Table 6. *Statistical calculations for triplet invariants estimated via equation* (I.7) *after the histogram-matching normalizing procedure (at* 4 Å *resolution)*

	Positive estimated triplets			Negative estimated triplets		
|ARG|	NR	%	⟨|Φ|°⟩	NR	%	⟨|Φ|°⟩
(*a*) TAQ (Hg derivative)						
0.4	22413	53	86	14520	51	91
0.8	21982	53	86	14103	51	91
1.2	12875	53	85	7088	52	92
1.6	5760	54	84	2731	52	92
2.0	2260	55	84	919	52	92
2.6	484	54	85	183	50	90
3.2	82	54	87	26	54	99
3.8	5	40	105	3	0	43
(*b*) FIS						
0.2	31019	55	84	18981	54	95
0.4	16948	56	84	7094	56	96
0.8	3232	59	80	923	57	97
1.2	657	59	80	141	55	96
1.6	148	56	82	22	64	106
2.0	27	59	81	1	100	138
(*c*) NOX						
0.2	25465	56	83	24535	55	96
0.4	24365	58	82	20432	56	97
0.8	6519	62	76	5211	61	103
1.2	1901	66	70	1457	67	110
1.6	593	70	67	473	70	114
2.0	190	76	56	146	80	124
2.6	33	79	54	21	95	143
3.2	4	100	33	3	100	152
(*d*) BPO (Au derivative)						
0.4	25072	69	68	24928	68	111
0.8	4936	77	56	3891	77	121
1.2	788	84	609	84	89	114
1.6	139	90	39	96	93	137
2.0	18	94	33	14	100	144
(*e*) BPO (Pt derivative)						
0.4	24937	71	65	24028	71	114
0.8	3809	81	53	3161	82	127
1.2	542	87	44	437	91	139
1.6	92	92	34	56	96	149
2.0	12	100	28	3	100	162

We note that:

(*a*) APP, NOX, M-FABP, TAQ(Pt) and BPO curves are too sharp. As a consequence, the $|\Delta|$'s are underestimated and reliable triplets are weakly discriminated from unreliable ones. This explains the anomalous triplet statistics of NOX and BPO shown in Tables 4(*c*), (*d*) and (*e*).

(*b*) FIS and TAQ(Hg) curves are markedly shifted towards the left. As a consequence, the ratio RPM will generally be smaller than its expected value and the percentage of negative triplets will be abnormally high. This explains the poor triplet statistics shown in Tables 4(*a*) and (*b*) for TAQ(Hg) and FIS.

(*c*) E2 and CARP curves are sufficiently close to the theoretical ones. It is therefore not surprising that, among the test structures, only E2 and CARP (see Table 5) were satisfactorily phased by the procedure described in papers II and III.

Table 7. *The phasing procedure is applied by using the histogram-matching normalizing procedure described in the text, at* 4 Å *resolution data*

Order of solution is the order of solution as ranked by CFOM. NPHAS is the number of phased reflexions, ERR is the average phase error calculated with respect to published phase values.

Structure code	Order of solution	NPHAS	ERR (weighted)
APP	2	143	38 (36)
CARP	3	391	44 (40)
E2	1	1750	35 (32)
M-FABP	1	580	58 (54)
BPO (Au)	1	2583	30 (23)
BPO (Pt)	1	2442	24 (19)
FIS	43	835	68 (60)
NOX	1	740	58 (42)
TAQ (Pt)	5	3867	71 (70)
TAQ (Hg)	–	–	–

The above observations suggest that $P(\Delta)$ may be conveniently used as a target distribution with which the experimental curves should comply. We do this according to the following procedure:

(1) the B_p value is found by the standard Wilson method using all the reflections up to native resolution.

(2) ΔB and $R_K = K_d/K_p$ are found by differential Wilson plot. Then, B_d and K_d are set to $B_d = B_p + \Delta B$ and $K_d = K_p R_K$.

(3) The scale factor K_d is suitably modified in order to satisfy the expected RPM at the chosen σ value for $\Delta_T = 0$.

(4) Histogram-matching techniques are (Zhang & Main, 1990) applied to transform the experimental curve into the $P(\Delta)$ distribution expected at the chosen σ value. Equation (I.7) is then applied to the Δ values so obtained for estimating triplet invariants.

It is instructive to compare triplet statistics obtained by the new procedure (see Table 6) with statistics shown in Table 4. It is immediately seen that systematic errors in the triplet sign estimation for FIS and TAQ(Hg) are avoided by the new normalizing procedure. Furthermore, triplet statistics of NOX and BPO are improved: the range of G values in Tables 6(*c*), (*d*) and (*e*) is quite reasonable and good triplets are more efficiently discriminated from unreliable ones.

The application of the phasing procedure to the new Δ's at 4 Å resolution data gives the results shown in Table 7. We note: (*a*) CARP and E2, for which a solution was found in Table 5, are again solved; (*b*) a satisfactory solution is now found for APP, M-FABP, BPO and NOX. It is worthwhile stressing the spectacular result obtained for BPO. A noisy solution is also found for FIS and TAQ(Pt) even if with the penalty of a large mean phase error.

The histogram-matching normalizing procedure at derivative resolution

It is useful to check if the normalizing procedure described in paper II might be successfully applied at derivative resolution to all our test proteins, and also to cases like FIS, NOX and TAQ (Pt derivative) for which only low-quality derivatives are available. We show in Figs. 9–12 the experimental curves $P(\Delta)$ at derivative resolution together with the theoretical curve calculated for the representative σ value. Comparison of Figs. 5–8 with Figs. 9–12 shows that: (*a*) the APP curve in Fig. 9 is shifted towards the right and the fit with the theoretical curve improves; on the contrary, the FIS curve does not remarkably change with resolution; (*b*) while E2 and

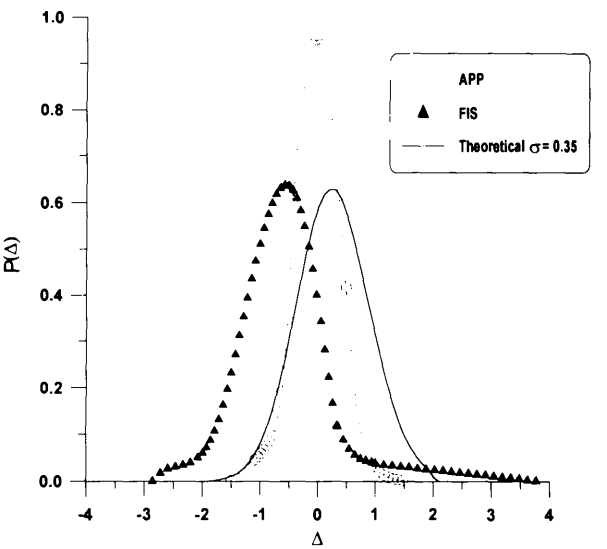

Fig. 9. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.35$ and corresponding experimental curves for APP and FIS obtained by the normalizing procedure described in paper II (derivative resolution data).

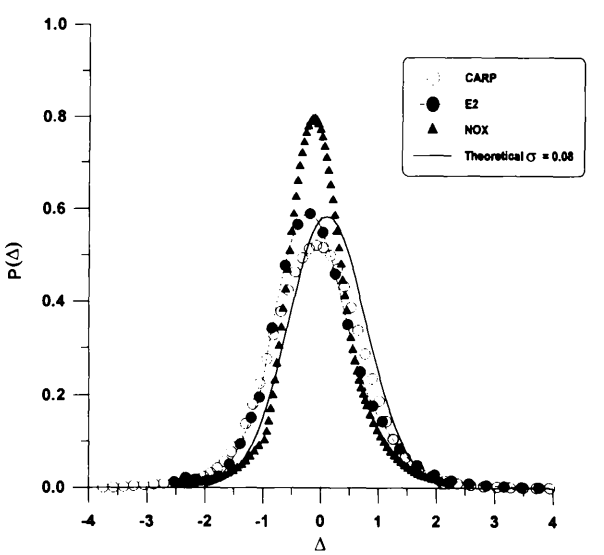

Fig. 10. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.08$ and corresponding experimental curves for CARP, E2 and NOX, obtained by the normalizing procedure described in paper II (derivative resolution data).

CARP do not change with resolution (they remain sufficiently good), the fits of NOX and BPO are remarkably better at derivative resolution; (*c*) the M-FABP experimental curve greatly improves at high resolution.

One can conclude that the normalizing procedure proposed in paper II improves as resolution increases. It is not a surprise then that the phasing process described in papers II and III works well at 3 Å or higher resolution and that NOX and BPO can be solved at derivative resolution, even if for NOX with the penalty of a large mean phase error (see Table 8). If the histogram-matching normalizing procedure is used, the phasing process produces the results shown in Table 8. For NOX, it should be noticed that the order of solution is 7. However, the three trials with order 1, 2, 3 are also solutions, even if with a larger phase error (73°, weighted error = 67°). By comparing the effects of the old normalizing procedure with those produced by the new one, we conclude that the new normalizing procedure is preferable.

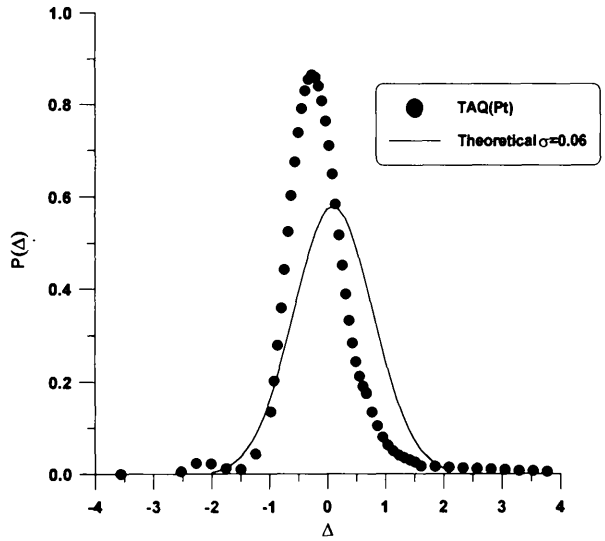

Fig. 11. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.06$ and corresponding experimental curves for TAQ (Pt derivative) obtained by the normalizing procedure described in paper II (derivative resolution data).

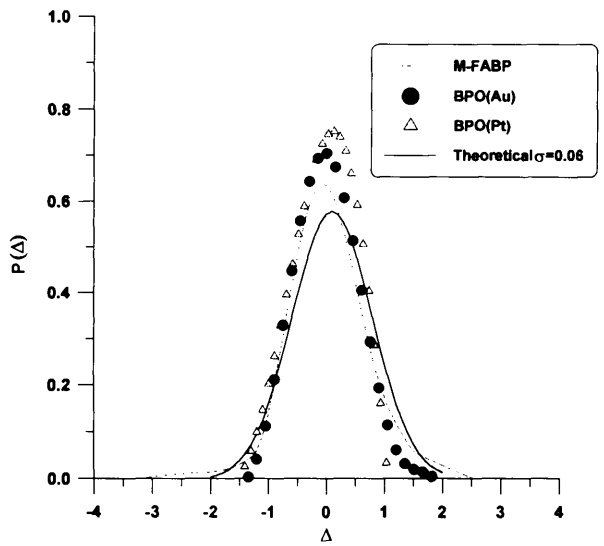

Fig. 12. $P(\Delta)$ distribution curve theoretically expected at $\sigma = 0.06$ and corresponding experimental curves for BPO (Au and Pt derivatives) and M-FABP obtained by the normalizing procedure described in paper II (derivative resolution data).

Table 8. *The phasing procedure is applied at derivative resolution data by: (a) using the normalizing procedure described in papers* II *and* III (*'old'*); *(b) the histogram-matching normalizing procedure described in the text* (*'new'*)

Order of solution is the order of solution as ranked by CFOM. NPHAS is the number of phased reflexions, ERR is the average phase error calculated with respect to published phase values. 100 trials have been calculated.

Structure code	Procedure	Order of solution	NPHAS	ERR (weighted)
APP	Old	3	810	46 (43)
	New	2	988	45 (41)
CARP	Old	2	2111	50 (46)
	New	2	2443	51 (48)
E2	Old	3	3218	40 (37)
	New	2	3662	41 (36)
M-FABP	Old	1	1231	50 (47)
	New	1	3330	54 (51)
BPO (Au)	Old	1	8307	28 (19)
	New	1	8343	28 (20)
BPO (Pt)	Old	1	–	–
	New	–	7908	34 (25)
FIS	Old	–	–	–
	New	92	1242	67 (69)
NOX	Old	4	1827	67 (60)
	New	7	1842	68 (61)
TAQ (Pt)	Old	–	–	–
	New	–	–	–

Discarding false solutions

In Tables 3–6 of paper II, we showed, for the four test structures, the figures of merit (FOM's) of the trial solutions with the highest values of CFOM. The solution with the highest value of CFOM was the correct solution for M-FABP, but that seemed to be the exception not the rule. Indeed, for CARP and E2, the solutions with highest value of CFOM were devoid of structural meaning, while for APP the correct solution had CFOM quite similar to that of a false solution. More efficient FOM's seem necessary for discriminating the correct from the false solutions. In their absence, it would be useful to be able to fix some criteria that could help to discard the false solutions with high values of CFOM. The scenario may be the following. Suppose that the phasing procedure has produced various trial solutions at the end of the phase-extension process described in paper III. They are ranked in order of CFOM. Then:

Table 9. *Heavy-atom positions from Fourier synthesis with coefficients* $(F_d - F_p)\exp(i\Phi_p)$ *for the highly ranked trial solutions (derivative resolution data)*

Structure code, space group	SET (CFOM)	Heavy-atom positions x	y	z	Height of the peaks
APP	1 (0.68)	0.50	0.29	0.50	369
*C*2	**2 (0.57)**	**0.75**	**0.45**	**0.23**	**250**
	3 (0.36)	0.50	0.31	0.33	243
CARP	1 (0.77)	0.00	0.00	0.00	112
*C*2	**2 (0.57)**	**0.76**	**0.17**	**0.09**	**192**
	3 (0.39)	0.47	0.31	0.38	133
E2	1 (1.0)	0.00	0.00	0.50	185
*F*432	**2 (0.96)**	**0.21**	**0.07**	**0.20**	**397**
	3 (0.55)	0.09	0.00	0.09	119
M-FABP	**1 (0.40)**	**0.89**	**0.06**	**0.74**	**648**
$P2_12_12_1$	2 (0.35)	0.09	0.15	0.59	670
BPO (Au)	**1 (0.89)**	**0.41**	**0.03**	**0.78**	**446**
$P2_13$		**0.78**	**0.11**	**0.81**	**320**
	2 (0.88)	**0.59**	**0.03**	**0.28**	**636**
		0.21	**0.11**	**0.31**	**541**
	3 (0.63)	0.15	0.15	0.15	410
		0.94	0.06	0.55	293
BPO (Pt)	**1 (0.85)**	**0.41**	**0.03**	**0.78**	**446**
$P2_13$		**0.78**	**0.11**	**0.81**	**320**
	2 (0.74)	0.04	0.04	0.04	1304
		0.11	0.11	0.11	246
NOX	**1 (0.77)**	**0.24**	**0.11**	**0.47**	**526**
$P4_12_12$	**2 (0.61)**	**0.76**	**0.11**	**0.27**	**537**
	7 (0.56)	**0.74**	**0.11**	**0.77**	**775**

(*a*) Difference Fourier syntheses with coefficients $(F_d - F_p)\exp(i\varphi_p)$ are calculated for the solutions with the highest values of CFOM. The maxima in the map should provide heavy-atom positions.

(*b*) Such parameters are refined according to the phase refinement process (Dickerson, Kendrew & Strandberg, 1961).

(*c*) If the refined positional parameters coincide with an allowed origin of the protein space group, then the trial solution is discarded from the set of reliable ones.

Steps (*a*), (*b*) and (*c*) are executed in sequence without user intervention.

Why should such a process work? Readers accustomed to direct phasing of small molecules know that in symmorphic space groups the so-called 'uranium solution' occurs quite frequently. It is marked by a high consistency of triplet phases, which are all close to zero. An observed Fourier synthesis would produce a huge maximum at an allowed origin. This type of false solution may be recognized and therefore discarded by special FOM's like the psi-zero and negative-quartet criteria. Since the psi-zero FOM described in paper II is not highly discriminating for macromolecules and the negative-quartet criterion is not among the used FOM's, the calculation of the difference Fourier synthesis is an efficient substitute for the specific FOM's. It is worthwhile emphasizing that a difference Fourier synthesis for proteins should not provide huge maxima at the allowed origins as for small molecules: since our phasing procedure uses a nearly equivalent number of positive and negative triplets, peak intensities in the maps corresponding to the 'uranium solutions' are similar to peak intensities corresponding to true heavy-atom positions.

In Table 9, we show, for each test structure and for the trial solutions highly ranked by CFOM, the heavy-atom positions as obtained after some cycles of Fourier–least-squares calculations. Data corresponding to the correct solution are in bold. If use is made of the information in Table 9, the correct solution is unambiguously recognized.

Concluding remarks

The phasing procedure described in papers II and III of this series has been improved in several aspects. First, a more robust normalizing procedure has been designed that makes explicit use of the distribution $P(\Delta)$. Histogram-matching procedures are used to obtain an optimal fit of the observed Δ distribution with the expected one. The new Δ's are statistically more meaningful and are able in most cases to overcome the disturbing effects provoked on the Wilson method for data up to 4 Å resolution by the presence of strong Debye effects.

A method is also suggested for discarding some false solutions provided by our multisolution technique. Since FOM's cannot safely work for isomorphous data where the signal is often comparable with the noise, an *a posteriori* check on the heavy-atom positions allows one to discard those trials that correspond to what are called 'uranium solutions' in the small-molecule direct-methods applications.

The authors are grateful to Drs H.-J. Hecht, W. Hol, N. Krauss, A. Mattevi, W. Saenger and G. Zanotti for having provided protein diffraction data and for useful discussions.

References

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

CASCARANO, G., GIACOVAZZO, C. & GUAGLIARDI, A. (1992). *Z. Kristallogr.* **200**, 63–71.

DICKERSON, R. E., KENDREW, J. C. & STRANDBERG, B. E. (1961). *Acta Cryst.* **14**, 1188–1195.

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

GIACOVAZZO, C., SILIQI, D. & RALPH, A. (1994). *Acta Cryst.* A**50**, 503–510.

GIACOVAZZO, C., SILIQI, D. & SPAGNA, R. (1994). *Acta Cryst.* A**50**, 609–621.

GIACOVAZZO, C., SILIQI, D. & ZANOTTI, G. (1995). *Acta Cryst.* A**51**, 177–188.

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

HALL, S. R. & SUBRAMANIAN, Y. (1982). *Acta Cryst.* A**38**, 590–598, 598–608.

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

HECHT, H.-J., ERDMANN, H., PARK, H.-J., SPRINZL, M. & SCHMID, R. D. (1995). In preparation.

HECHT, H.-J., ERDMANN, H., PARK, H.-J., SPRINZL, M., SCHMID, R. D. & SCHOMBURG, D. (1993). *Acta Cryst.* A**49**, 86.

HECHT, H.-J., SOBEK, H., HAAG, T., PFEIFER, O. & VAN PEE, K. H. (1994). *Nature (London) Struct. Biol.* **1**, 532–537.
KOSTREWA, D., GRANZIN, J., STOCK, D., CHOE, H.-W., LABAHN, J. & SAENGER, W. (1992). *J. Mol. Biol.* **226**, 209–226.
KRETSINGER, R. H. & NOCKOLDS, C. E. (1973). *J. Biol. Chem.* **248**, 3313–3326.
LABAHN, J., GRANZIN, J., SCHLUCKEBIER, G., ROBINSON, D. P., JACK, W. E., SCHILDKRAUT, I. & SAENGER, W. (1995). *Proc. Natl Acad. Sci. USA*. In press.
MATTEVI, A., OBMOLOVA, G., SCHULZE, E., KALK, K. H., WESTPHAL, A. H., DE KOK, A. & HOL, W. G. J. (1992). *Science*, **255**, 1544–1550.
RICHARDSON, J. S. & RICHARDSON, D. C. (1985). *Methods in Enzymology*, Vol. 115B, edited by H. W. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF, pp. 189–206. Orlando: Academic Press.
SUBRAMANIAN, V. & HALL, S. R. (1982). *Acta Cryst.* A**38**, 577–590.
ZANOTTI, G., SCAPIN, G., SPADON, P., VEERKAMP, J. H. & SACCHETTINI, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.
ZHANG, K. Y. J. & MAIN, P. (1990). *Acta Cryst.* A**46**, 41–46.

Acta Cryst. (1995). A**51**, 820–825

The Joint Probability Distribution of Any Set of Phases Given Any Set of Diffraction Magnitudes. IV. The Active Use of Psi-Zero Triplets

BY G. CASCARANO AND C. GIACOVAZZO

Istituto di Ricerca per lo Sviluppo di Metodologie Cristallografiche CNR, c/o Dipartimento Geomineralogico, Campus Universitario, 70124 Bari, Italy

(*Received* 5 *December* 1994; *accepted* 14 *March* 1995)

Abstract

In some recent papers [Giacovazzo, Burla & Cascarano (1992). *Acta Cryst.* A**48**, 901–906; Burla, Cascarano & Giacovazzo (1992). *Acta Cryst.* A**48**, 906–912; Cascarano, Giacovazzo, Moliterni & Polidori (1994). *Acta Cryst.* A**50**, 22–27], the method of the joint probability distribution of structure factors has been used to define a function that is frequently a maximum for the correct structure. Such a function was the basis for a modified tangent formula using $P10$ (negative and positive) triplet estimates and negative quartet estimates, which proved more efficient than the classical tangent formula of Karle & Hauptman [*Acta Cryst.* (1956), **9**, 635–651]. The method is here combined with a recent formulation [Giacovazzo (1993). *Z. Kristallogr.* **206**, 161–171], which suggests the supplementary active use in the phasing process of psi-zero triplets. Experimental tests prove the higher efficiency of the method and justify the default active use of psi-zero relationships in *SIR*92.

Symbols and notation

Symbols and notation are the same as in the following papers; Giacovazzo, Burla & Cascarano (1992); Burla, Cascarano & Giacovazzo (1992); Cascarano, Giacovazzo, Moliterni & Polidori (1994); from now on these are referred to as papers I, II and III, respectively.

Introduction

In papers I and II of this series, the conditional joint probability distribution of n phases given p ($p \geq n$) moduli was studied. Large values of n and p are allowed; *e.g.* n may be the number of strong reflections that are usually phased by a modern direct procedure and p may be the number of measured reflections. The resulting distribution is of exponential type and contains triplet and quartet contributions. Contrary to any expectations, the distribution is not maximized by the correct set of phases as one would expect for sufficiently large values of n and p. Accordingly, the combined use of triplets and quartets proved of limited usefulness for practical applications.

In paper III, the failure of the distribution was ascribed to the strong correlation between triplets and positive estimated quartets. A modified expansion of the distribution was then proposed that neglects the contribution of the positive estimated quartets and retains terms arising from triplets and negative quartets only. The distribution is often maximized by the correct solution. Accordingly, an efficient tangent formula was described actively using triplet [estimated positive or negative by the $P10$ formula (Cascarano, Giacovazzo, Camalli, Spagna, Burla, Nunzi & Polidori, 1984)] and negative quartet relationships.

The main guidelines of the three papers may be summarized as follows: if a suitable function may be found that is maximum for the correct solution, then a tangent formula may be designed that is expected to be more efficient than the traditional tangent formula of Karle & Hauptman (1956). The function to maximize is in practice a figure of merit (FOM), a tool for recognizing the correct solution among numerous trial solutions.

This idea was first formulated in a recent paper by Giacovazzo (1993) where a new method is proposed that actively uses the information contained in the psi-zero triplets in order to drive phases towards the correct

ISSN 0108-7673