Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

An improved phase-extension procedure for isomorphous replacement phases

Yanina Vekhter[a]* and Russ Miller[b,c]

[a]Department of Physics, State University of New York at Buffalo, Buffalo, NY 14260, USA, [b]Hauptman-Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203, USA, and [c]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260, USA

Correspondence e-mail: dvekhter@acsu.buffalo.edu

Received 12 September 2000
Accepted 4 April 2001

A new phase-extension procedure has been applied to isomorphous replacement data and shown to yield improved phases and maps compared with standard solvent flattening operating on a full set of centroid phases. In this procedure, a starting subset of core phases is selected based on the sharpness of the phase-probability curves. Phase extension using solvent flattening as the density-modification procedure is then carried out, gradually adding additional phases. In tests with known protein structures, the mean phase errors for the output expanded phase sets were reduced by 3–9° and the corresponding map correlation coefficients were increased by 0.05–0.18 relative to phase sets from standard solvent-flattening procedures. With SIR data, the lowest final mean phase errors were approximately 58° and the corresponding map correlation coefficients were in the range 0.53–0.68.

1. Introduction

In the isomorphous replacement method, as implemented by Blow & Crick (1959), initial estimates for protein phases are assigned based on phase-probability curves. The probability of a phase $P(\alpha)$ is evaluated as

$$P(\alpha) = N[-\varepsilon(\alpha)^2/2E^2], \qquad (1)$$

where $\varepsilon(\alpha) = |F_{\rm PH_{calc}}| - |F_{\rm PH_{obs}}|$ is the discrepancy between the magnitudes of the observed and calculated structure factors for a heavy-atom derivative, $|F_{\rm PH_{calc}}| = F_P + F_H$ is found as the sum of protein (F_P) and heavy-atom (F_H) contributions, N is a normalization factor and E^2 is the mean-square value of ε. For SIR data, the phase-probability curves are bimodal for non-centric reflections (Bokhoven *et al.*, 1951; Blow & Crick, 1959). The maxima correspond to two possible phase values, one of which is true and the other false. This phase ambiguity is generally resolved by additional data based either on the anomalous signal from the heavy atoms measured at a single (SIRAS; Rossmann & Blow, 1963; North, 1965; Matthews, 1966) or at multiple wavelengths (MAD; Okaya & Pepinsky, 1956; Hendrickson, 1985; Fourme & Hendrickson, 1990) or on the scattering from multiple isomorphous heavy-atom derivatives (MIR; Green *et al.*, 1954). When information from various sources is combined, the resultant probability for a phase $P(\alpha)$ is found as the product of individual probabilities. The so-called 'best' or centroid maps are calculated based on centroid phases, $\alpha_{\rm cent}$, which correspond to the center of gravity of the phase-probability distribution and the weighted Fourier coefficients $|F_{\rm best}(H)|$,

$$|F_{\rm best}(H)| = w|F_P| = |F_P|\frac{\int_0^{2\pi}\exp(i\alpha)P(\alpha)\,{\rm d}\alpha}{\int_0^{2\pi}P(\alpha)\,{\rm d}\alpha}. \qquad (2)$$

Centroid maps computed using isomorphous replacement phases may be improved by solvent flattening (Wang, 1985; Leslie, 1987) or other forms of density modification (Rossmann & Blow, 1963; Lunin, 1988; Zhang & Main, 1990; Baker *et al.*, 1993). In these procedures, physical and chemical information about a structure is imposed as constraints on the electron density. In the solvent-flattening procedure, for example, the molecular boundary is located, the density in the bulk-solvent region is set to a constant low value and negative density in the protein region is truncated. In the SIR case, the *full* set of centroid phases typically has a mean phase error in the range 60–70° even after density modification. Therefore, it is usually impossible to build a starting model for a protein structure based on a centroid map. The major contribution to SIR centroid mean phase error arises from non-centric phases with bimodal phase probability curves having the two maxima further apart than 90°, as well as both centric and non-centric phases that do not satisfy the closure condition $|F_{\rm PH} - F_P| \cong F_H$. For such centric phases, the two restricted phase values have almost equal probability. The corresponding non-centric phases have flat unimodal phase probability curves and are biased by the heavy-atom phases. The mean phase error for a subset of SIR phases is in the range 35–45° when these two groups of phases are excluded. Since the accuracy of an individual phase

determination depends on the sharpness of the phase-probability curve near the maximum value, a threshold cutoff value P_{cut} can be applied to the probability-curve maximum to select a subset of phases with lower than average mean phase error. The purpose of the present study was to determine whether a 'better start' using a more accurate subset of phases (the core) improves the final full set of phases following solvent flattening.

2. Materials and methods

Solvent flattening was applied to SIR, SIRAS and MIR data for the six known protein structures described in Table 1. Both the standard procedure implemented in the computer program *PHASES* (Furey & Swaminathan, 1997) as well as phase extension from a core set as implemented in a modified version of this program were used.

Table 1
Test data sets: porphobilinogen deaminase (PBGD), actinoxanthin (ACT), porcine elastase (ELA), killer toxin KP6-α (KP6), macromomycin (MCM) and rabbit liver fructose-1,6-diphosphatase (FDP).

Test structure	Space group	PDB code	No. of residues	Solvent content (%)	Heavy-atom derivatives	Native resol. (Å)	Deriv. resol. (Å)	References
PBGD	$P2_12_12$	1pda	314	50	UO_2F_2	1.9	3.0	Louie *et al.* (1992)
					K_2PtCl_4		3.0	
ACT	$P2_12_12_1$	1acx	108	40	UO_2F_2	2.0	3.0	Pletnev *et al.* (1982)
ELA	$P2_12_12_1$	1lvy	245	40	Kr	1.9	1.9	Schiltz *et al.* (1997)
KP6	$P6322$	1kp6	79	50	Se Met	2.2	3.0	Li *et al.* (1999)
					Pt_2NO_3		3.0	
MCM	$P2_1$	2mcm	115	36	$K_2Pt(NO_2)_4$	1.5	2.5	Van Roey & Beerman (1989)
FDP	$I222$	1bk4	337	54	$K_3UO_2F_5$	2.3	3.0	Weeks *et al.* (1999)
					$KAu(CN)_2$		3.5	

The phase-extension procedure consisted of the following steps.

(i) Core phases were selected by first excluding non-centrics with two maxima separated by more than 90°. Threshold cutoff values P_{cut} were then applied to the remaining phases. P_{cut} was set to 0.03 for non-centrics and to 0.7 for centrics in the SIR and SIRAS cases and to 0.05 for non-centrics and 0.8 for centrics in the MIR and MIRAS cases. The rationales for choosing these values of P_{cut} are presented in §3.

(ii) A series of phase-extension steps was then performed in such a way that the lowest resolution unphased reflections, equal to 2% of the core phases, were added at every step.

(iii) During each extension step, four cycles of solvent flattening were performed with either one or two masks. The first mask was based on the expanded phase set from the previous step and the second mask (when used) included the full set of phases. For SIRAS and MIR data, as well as actinoxanthin and KP6α (Se derivative) SIR data, solvent flattening was performed with only one mask.

(iv) The initial core phases were held fixed at their centroid values and their weights (W) were set equal to unity until extension to 5 Å was complete. These reflections were allowed to refine during extension beyond 5 Å.

(v) During any phase-extension step, the phases α_{calc} and magnitudes $|F_{calc}|$ were computed by Fourier transformation of the solvent-flattened map and the phase probabilities were generated according to Sim (1959) and Bricogne (1974) as

$$P(\alpha) = N' \exp\left[-\frac{(F_{obs} - F_{calc})^2}{\Sigma_q}\right] = N \exp[X \cos(\alpha - \alpha_c)], \quad (3)$$

where $X = 2F_{obs}F_{calc}/\Sigma_q$ and Σ_q was evaluated based on the discrepancies between F_{obs} and F_{calc} averaged over resolution shells so that $\Sigma_q = \langle|F_{obs}^2 - F_{calc}^2|\rangle$. The phase probabilities for the output phases were computed as the product of the probabilities of the starting centroid phases and the phases from the final solvent-flattened maps. Fourier coefficients with the usual weights,

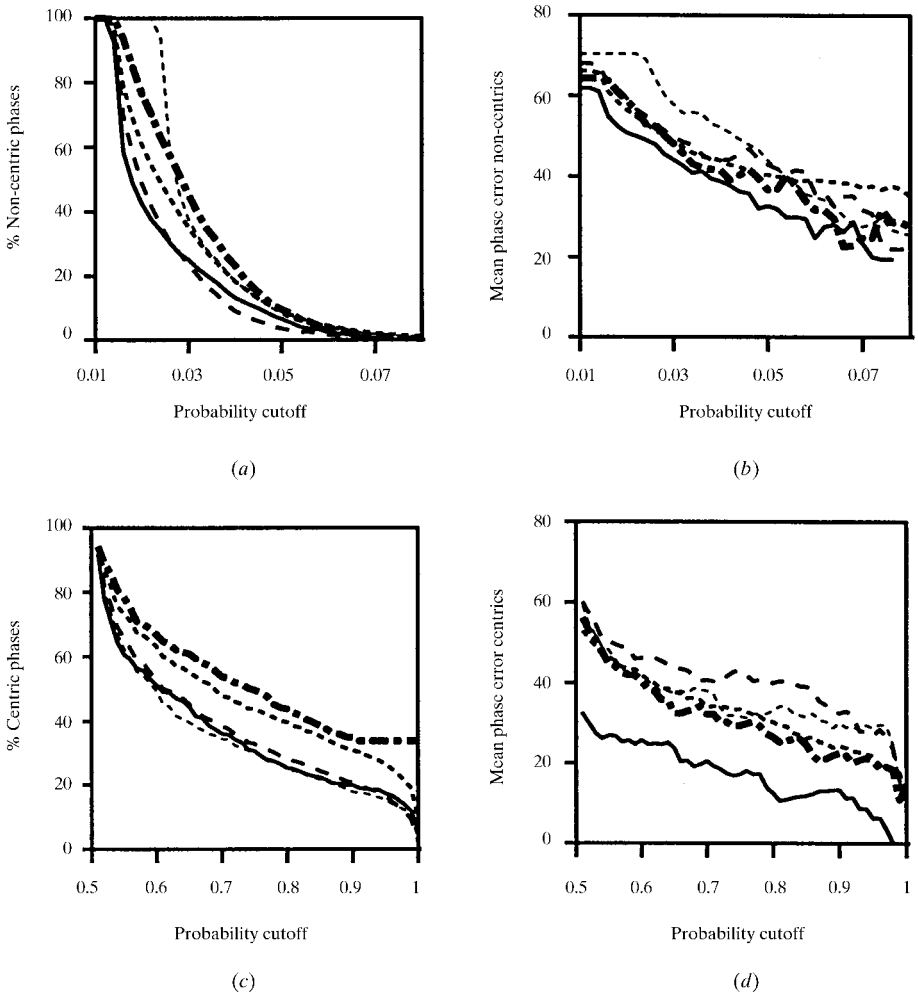

Figure 1
Percentages of reflections remaining as core phases (a, c) and the corresponding mean phase errors (b, d) for non-centric and centric data, respectively, as a function of probability cutoff (P_{cut}) for the SIR test structures (continous line, ACT; long dashes, PBGD; short dashes, FDP $K_3UO_2F_5$; short bold dashes, ELA; dot–dash lines, KP6 Se).

Table 2
Percentage of the core phases and mean phase errors for the core; mean phase errors and map correlation coefficients for centroid phases and after standard solvent flattening and phase-extension procedures.

		Centroid phases (full set)			Standard solvent-flattening procedure		Phase extension	
Test structure	Core phases (%)	$\langle\Delta\varphi\rangle$	$\langle\Delta\varphi\rangle$	CC	$\langle\Delta\varphi\rangle$	CC	$\langle\Delta\varphi\rangle$	CC
ACT (SIR)	16	35.3	66.6	0.42	66.4	0.50	57.4	0.68
ELA (SIR)	23	46.6	70.1	0.42	59.0	0.61	59.3	0.62
KP6 [SIR(Pt)]	19	45.6	72.0	0.39	71.8	0.41	70.8	0.44
KP6 [SIR(Se)]	28	43.2	68.7	0.41	61.4	0.48	58.2	0.53
KP6 [SIRAS(Se)]	23	39.4	66.0	0.44	60.4	0.48	60.2	0.55
KP6 (MIR)	23	36.6	59.4	0.50	53.7	0.58	52.2	0.62
KP6 (MIRAS)	40	39.6	56.6	0.54	50.8	0.62	47.0	0.68
MCM (SIRAS)	53	32.7	48.5	0.66	37.0	0.82	34.0	0.86
PBGD [SIR(UO_2F_2)]	18	47.4	73.2	0.31	72.6	0.33	67.9	0.44
PBGD (MIR)	28	44.9	66.8	0.42	57.4	0.60	55.4	0.66
FDP (SIRAS)	40	57.0	70.8	0.38	63.8	0.47	63.0	0.51
FDP (MIR)	51	51.7	62.9	0.45	56.1	0.57	53.4	0.63

$$W = \left(\left\{ \frac{\sum [P(\alpha)\sin(\alpha)]}{\sum P(\alpha)} \right\}^2 + \left\{ \frac{\sum [P(\alpha)\cos(\alpha)]}{\sum P(\alpha)} \right\}^2 \right)^{1/2}, \quad (4)$$

employed in the *PHASES* program, were used to compute maps.

The extension procedure was completed after all reflections were phased. The unweighted mean phase errors,

$$\langle\Delta\varphi\rangle = \frac{1}{N}\sum_i |\varphi(i) - \varphi_{\rm model}(i)|, \quad (5)$$

and map correlation coefficients,

$$\mathrm{CC} = \frac{\{\sum[\rho(r) - \langle\rho(r)\rangle][\rho_{\rm model}(r) - \langle\rho_{\rm model}(r)\rangle]\}}{\left(\{\sum[\rho(r) - \langle\rho(r)\rangle]^2 \sum[\rho_{\rm model}(r) - \langle\rho_{\rm model}(r)\rangle]^2\}^{1/2}\right)}, \quad (6)$$

were calculated based on the known structure models and the extended phase sets.

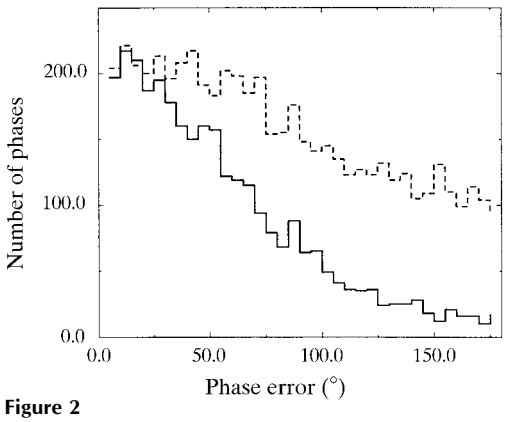

Figure 2
Phase-error histograms for the porcine elastase non-centric phases that have two probability maxima separated by less than 90°. Phases with probability maxima ≤ 0.03 (dashed line) and phases with probability maxima >0.03 (full line).

3. Results and discussion

The results of the phase-extension procedure described above, as well as the results of the standard solvent-flattening procedure implemented in the computer program *PHASES* (Furey & Swaminathan, 1997), are summarized in Table 2. With SIR data, the most striking effects of slow phase extension compared with the standard solvent-flattening procedure occurred for the actinoxanthin structure, where the final value of the mean phase error improved by 9° and the map correlation coefficient improved by 0.18. For SIRAS and MIR data, the mean phase errors after the phase-extension procedure were reduced by about 3° for macromomycin (SIRAS), KP6-α (MIRAS) and Fru-1,6-Pase (MIRAS) and the corresponding map correlation coefficients improved by 0.04–0.06. Additional comparison tests were run with the *DM* program (Collaborative Computational Project, Number 4, 1994) on the three test structures [ACT, PBGD and KP6-α(Se)] that showed the greatest improvement with the phase-extension procedure SIR data. With *DM*, the output improved for the PBGD structure compared with *PHASES* (the mean phase error and map correlation coefficient improved by 3.1° and 0.09 and were equal to 69.5° and 0.42; for other structures the results were consistent with *PHASES*).

The phase-extension procedure yielded an improved output based on a more accurate starting subset of phases. To select the core phases, a higher cutoff should be applied to centric phases than to non-centrics, since the phase probabilities for non-centrics are evaluated throughout the interval (0, 360°), whereas the probabilities for centrics are evaluated just for two phase values that are 180° apart and the normalized probability at the maximum is much higher for centrics. As illustrated in Fig. 1 for SIR data, as the probability cutoff threshold ($P_{\rm cut}$) increases both the percentage of the phases in the core and the mean phase errors for these phases decrease. The success of phase extension depends on the choice of $P_{\rm cut}$ values. For the test SIR cases, the phase-extension procedure yields the best results with the threshold values for $P_{\rm cut}$ selected at 0.03 for non-centrics and at 0.7 for centrics. These $P_{\rm cut}$ values are suggested to be used as defaults. As illustrated in Fig. 2 with the phase-error histogram of the porcine elastase, for non-centric phases with the probability maxima above 0.03 percentage of phases with mean errors exceeding 60° is significantly lower compared with the phases with $P_{\rm cut} \leq 0.03$. As shown in Fig. 1, at $P_{\rm cut}$ values of 0.7 and 0.03 approximately 30% of the phases were included in the core and the mean phase errors were in the ranges 20–43° for centrics and 40–60° for non-centrics. In MIR and MIRAS cases, the phase ambiguity is resolved for a higher percentage of phases compared with SIR. Therefore, even though higher thresholds ($P_{\rm cut} \geq 0.05$ for non-centrics and $P_{\rm cut} \geq 0.8$ for centrics) were applied, about 50% of the phases (with corresponding mean phase errors of ~50°) were included in the core.

4. Conclusions

These studies have demonstrated that the new phase-extension procedure is more powerful for improving isomorphous replacement phases and maps than standard solvent flattening. The thresholds ($P_{\rm cut}$) for centric and non-centric phases that should be applied to select a starting subset of phases having mean phase error in the range of ~35–50° have been found. The mean phase error for the output phases after the phase-extension procedure is correlated with the mean phase error for the core phases. With SIR data and a 'better start' based on core phases having a mean error of ~35°, the phase ambiguity is shown to be correctly resolved for approximately 70% of the actinoxanthin phases (on the other hand, with standard solvent flattening operating on a full set of phases, the phase ambiguity is resolved for less than 60% of the phases). We can conclude, therefore, that this phase-extension procedure may be used to resolve the phase ambiguity for some SIR data sets or to obtain improved phases and maps for

SIRAS and MIR cases. The combination of phase extension with other density-modification procedures and additional constraints applied in the protein region should be investigated and is expected to yield improved results.

We would like to acknowledge Dr D. Ghosh for bringing this problem to our attention. We thank Drs I. J. Tickle, A. P. Kuzin, E. de La Fortelle, P. Van Roey, C. Weeks and N. Li for kindly supplying reflection data and refined structure models. We would like to express our appreciation to Melda Tugac and Gloria J. Del Bel for preparing the figures. In particular, we wish to thank Dr Charles Weeks for being a constant source of encouragement and for stimulating discussions. This research was supported by NIH grant GM-46733 and NSF grant ACI-9721373.

References

Baker, D., Bystroff, C., Fleterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–448.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Bokhoven, C., Schoone, J. C. & Bijvoet, J. M. (1951). *Acta Cryst.* **4**, 275–280.
Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. A*, **225**, 287–307.
Fourme, R. & Hendrickson, W. A. (1990). *Sychrotron Radiation and Biophysics*, edited by S. S. Hasnain, pp. 156–175. Chichester: Horwood.
Furey, W. & Swaminathan, S. (1997). *Methods Enzymol.* **277**, 590–620.
Hendrickson, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11–21.
Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.
Li, N., Erman, M., Pangborn, W., Duax, W. L., Park, C. M., Bruenn, J. & Ghosh, D. (1999). *J. Biol. Chem.* **274**, 20425–20431.
Louie, G. V., Brownlie, P. D., Lambert, R., Cooper, J. B., Blundell, T. L., Wood, S. P., Warren, M. J., Woodcock, S. C. & Jordan, P. M. (1992). *Nature (London)*, **359**, 33–39.
Lunin, V. Y. (1988). *Acta Cryst.* A**44**, 144–150.
Mathews, B. W. (1966). *Acta Cryst.* **20**, 82–86.
North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.
Okaya, Y. & Pepinsky, R. (1956). *Phys. Rev.* **103**, 1645.
Pletnev, V., Kuzin, A., Trakhanov, S. & Popovich, V. (1982). *Chemistry of Peptides and Proteins*, edited by W. Voelter, E. Wuensch, J. Ovchinnikov & V. Ivanov, Vol. 1, pp. 429–433. Berlin: Walter de Gruyter.
Rossmann, M. G. & Blow, D. M. (1963). *Acta Cryst.* **16**, 39–45.
Schiltz, M., Shepard, W., Fourme, R., Prange, T., de La Fortelle, E. & Bricogne, G. (1997). *Acta Cryst.* D**53**, 78–92.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
Van Roey, P. & Beerman, T. A. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 6587–6590.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Weeks, C. M., Roszak, A. W., Erman, M., Kaiser, R., Jornvall, H. & Ghosh, D. (1999). *Acta Cryst.* D**55**, 93–102.
Zhang, K. Y. D. & Main, P. (1990). *Acta Cryst.* A**46**, 41–46.