Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Combining direct methods with isomorphous replacement or anomalous scattering data. VIII. Phasing experimental SIR data with the replacing atoms in a centrosymmetric arrangement

**Yu-Dong Liu,[a] Yuan-Xin Gu,[a] Chao-De Zheng,[a] Q. Hao[a,b] and Hai-Fu Fan[a]***

[a]Institute of Physics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China, and [b]School of Applied Sciences, De Montfort University, Leicester LE1 9BH, England

Correspondence e-mail: fan@aphy.iphy.ac.cn

Received 12 September 1998
Accepted 15 December 1998

A multisolution direct method has been proposed to resolve the phase ambiguity intrinsic in single isomorphous replacement data of proteins with the replacing atoms in a centrosymmetric arrangement. The phase ambiguity of each reflection is in fact a 'sign ambiguity' of the phase difference between the phase of the native protein and that of the replacing atoms, *i.e.* $\pm |\Delta\varphi| = \varphi - \varphi'$. The $P_+$ probability formula can be used to derive the signs. The multisolution phasing procedure is initiated using random starting values of $P_+$. A cluster analysis is used instead of figures of merit to find the correct solution. The direct-method phases can be further improved by density-modification techniques. The method was tested with the experimental SIR data at 2 Å resolution from a known protein aPP; satisfactory results were obtained.

## 1. Introduction

It is well known in protein crystallography that phase ambiguity is intrinsic in single isomorphous replacement (SIR) data. The phase of a reflection from the native protein can be expressed as

$$\varphi_{\mathbf{H}} = \varphi'_{\mathbf{H}} \pm |\Delta\varphi_{\mathbf{H}}|, \quad (1)$$

where $\varphi_{\mathbf{H}}$ is the phase of the native protein for a reflection with the reciprocal vector equal to $\mathbf{H}$ and $\varphi'_{\mathbf{H}}$ is the phase of the replacing atoms, while $\Delta\varphi_{\mathbf{H}}$ is the difference between the phase of the native protein and that of the replacing atoms. Given a set of SIR data, it is not difficult to find the positions of the replacing atoms and then to derive $\varphi'_{\mathbf{H}}$ and $|\Delta\varphi_{\mathbf{H}}|$. The remaining problem is to find the sign of $\Delta\varphi_{\mathbf{H}}$, *i.e.* to break the phase ambiguity. A number of procedures have been proposed to solve this problem. These include the solvent-flattening method (Wang, 1981, 1985) and various kinds of direct methods (Coulter, 1965; Fan, 1965; Karle, 1966; Hendrickson, 1971; Hauptman, 1982; Fan & Gu, 1985; Fortier *et al.*, 1985; Klop *et al.*, 1987; Giacovazzo *et al.*, 1988; Kyriakidis *et al.*, 1993). In practice, the solvent-flattening method has been the most successful; however, it is impossible even in theory to resolve the phase ambiguity when the replacing atoms are in a centrosymmetric arrangement. On the other hand, direct methods are capable of breaking the phase ambiguity under these circumstances, as has been shown by Yao & Fan (1985) using a set of error-free SIR data. Nevertheless, until now no successful test using experimental data below atomic resolution has been reported. In this paper, we will describe a test with the experimental SIR data at 2 Å resolution from the known protein aPP, in which the replacing atoms are in a centrosymmetric arrangement.

## 2. Phasing

According to Fan & Gu (1985), the SIR phase ambiguity can be broken using the $P_+$ formula

$$P_+(\Delta\varphi_{\mathbf{H}}) = 0.5 + 0.5\Big(\tanh\Big\{\sin(|\Delta\varphi_{\mathbf{H}}|) \times \Big[\sum_{\mathbf{H}'} m_{\mathbf{H}'} m_{\mathbf{H}-\mathbf{H}'} \kappa_{\mathbf{H},\mathbf{H}'} \times \sin(\Phi'_3 + \Delta\varphi_{\mathbf{H}',\mathrm{best}} + \Delta\varphi_{\mathbf{H}-\mathbf{H}',\mathrm{best}})\Big]\Big\}\Big), \quad (2)$$

where

$$m_{\mathbf{H}} = \exp(-\sigma_{\mathbf{H}}^2/2)\{[2(P_+ - 0.5)^2 + 0.5] \times [1 - \cos(2\Delta\varphi_{\mathbf{H}})] + \cos(2\Delta\varphi_{\mathbf{H}})\}^{1/2}, \quad (3)$$

$$\tan(\Delta\varphi_{\mathbf{H},\mathrm{best}}) = 2(P_+ - 0.5)\sin|\Delta\varphi_{\mathbf{H}}|/\cos\Delta\varphi_{\mathbf{H}}, \quad (4)$$

$$\kappa_{\mathbf{H},\mathbf{H}'} = 2\sigma_3/\sigma_2^{3/2}|E_{-\mathbf{H}}E_{\mathbf{H}'}E_{\mathbf{H}-\mathbf{H}'}|, \quad \sigma_n = \sum_j Z_j^n, \quad (5)$$

and

$$\Phi'_3 = \varphi'_{-\mathbf{H}} + \varphi'_{\mathbf{H}'} + \varphi'_{\mathbf{H}-\mathbf{H}'}. \quad (6)$$

The factor $\exp(-\sigma_{\mathbf{H}}^2/2)$ in (3) is related to the 'lack-of-closure error' (Blow & Crick, 1959). To break the phase ambiguity, initial values should be assigned to $P_+$ for each reflection. (3) and (4) are then used to calculate values of $m_{\mathbf{H}}$ and $\Delta\varphi_{\mathbf{H},\mathrm{best}}$. The results are substituted into (2) to obtain a new set of $P_+$ values. The procedure can be made iterative. Finally, values of $\Delta\varphi_{\mathbf{H},\mathrm{best}}$ are converted to $\varphi_{\mathbf{H},\mathrm{best}}$ using (7),

$$\varphi_{\mathbf{H},\mathrm{best}} = \varphi'_{\mathbf{H}} + \Delta\varphi_{\mathbf{H},\mathrm{best}}. \quad (7)$$

In principle, there are two different ways to assign initial values to $P_+$. One is to assign all reflections to have an initial $P_+$ of 0.5. This implies a single-solution phasing procedure. The other is to assign random values to $P_+$ and implement a multisolution procedure. In the case that the replacing atoms are in a centrosymmetric arrangement, $\Phi'_3$ in (6) will always be 0 or $\pi$. Then, according to (4) and (2), an initial $P_+$ of 0.5 for all reflections will have all the newly calculated $P_+$ remaining unchanged. Hence, a multisolution phasing procedure is necessary for breaking the SIR phase ambiguity when the replacing atoms are in a centrosymmetric arrangement. In the test of Yao & Fan (1985), random values between 0.4 and 0.6 were used as the initial values of $P_+$. In the present test, however, we found that better results could be obtained by using random values between 0 and 1 for the initial $P_+$ values. A crucial step in multisolution phasing is to find an appropriate figure of merit to pick out the correct solution; however, this is not always easy, especially when the diffraction data available is below atomic resolution. In the present test, a cluster-analysis procedure was used instead of figures of merit.

## 3. Cluster analysis

Cluster analysis was introduced by Lunin *et al.* (1990) to determine the molecular envelope of proteins using very low resolution data. The philosophy of cluster analysis is that, among the phase sets resulting from a sufficiently large number of random trials, there should be a considerable number of sets close to the true solution, while others are randomly distributed and differ significantly from each other. Hence, we can find the cluster(s) that are closest to the true solution by grouping the phase sets as follows. The average phase difference $\langle \mathrm{dif}(\varphi)\rangle$ is calculated for every pair of phase sets according to

$$\langle \mathrm{dif}(\varphi)\rangle = \sum_{\mathbf{H}} F_{\mathbf{H}}[(\varphi_{\mathbf{H}})_1 - (\varphi_{\mathbf{H}})_2]/\sum_{\mathbf{H}} F_{\mathbf{H}}. \quad (8)$$

Then, for each phase set, the number of sets around it with a $\langle \mathrm{dif}(\varphi)\rangle$ of less than a given value, say 2°, is counted and the phase sets are arranged in descending order of these numbers. [In practice, this value is adjusted so that 76.2% (the area within the 'half-height width' of a Gaussian distribution) of the total phase sets are included in clusters.] The top phase set and all other sets having a $\langle \mathrm{dif}(\varphi)\rangle$ against the top set of less than 2° are grouped as the first cluster. The next phase set and those with a $\langle \mathrm{dif}(\varphi)\rangle$ against the second set of less than 2° are grouped as the second cluster and so on. Within each cluster, an average phase set may be obtained according to the following equation,

$$[m_{\mathbf{H}}\exp(i\varphi_{\mathbf{H},\mathrm{best}})]_{\mathrm{average}} = \sum_j^N m_{\mathbf{H}}\exp(i\varphi_{\mathbf{H},\mathrm{best}})]_j/N, \quad (9)$$

where $N$ is the number of phase sets in the cluster. The average phase sets from the largest clusters are taken as the most probable solutions. In the SIR case with the replacing atoms in a centrosymmetric arrangement, we can expect to find two largest clusters which correspond to the two enantiomorphs. If the unit-cell origin is chosen at the inversion centre of the replacing atoms, the phases from one enantiomorph will be the negative of those from the other. Hence, if two clusters can be identified as corresponding to solutions which are enantiomorphs, it may be useful to average the phases (after changing the signs where appropriate).

## 4. Test

Data used in the present test were from the native and the mercury derivative of a small protein, aPP (avian pancreatic polypeptide), at 2.0 Å resolution (Blundell *et al.*, 1981). The crystals belong to space group $C2$ with unit-cell parameters $a$ = 34.18, $b$ = 32.92, $c$ = 28.44 Å, $\beta$ = 105.3°. There is one molecule with 36 amino-acid residues in the asymmetric unit. The program used to break the phase ambiguity is a new version of the program *OASIS*. Apart from a number of minor modifications, the new version differs from the old one (Hao *et al.*, 1996) by the inclusion of multisolution phasing and cluster analysis. Details of the program will be described in a separate paper.

In the present test, 1000 random trials were implemented. Five iterative cycles were calculated for each trial. The cluster analysis yielded two large clusters amongst many smaller ones. Inspection of the phase difference between these two clusters

**Table 1**
Results on breaking the phase ambiguity in the SIR data from the known protein aPP.

Phase errors and map correlation coefficients were calculated against the structure model (Blundell *et al.*, 1981) excluding the crystallized water molecules. All 2106 observed reflections at 2 Å resolution from the native protein were used in the calculation. The SIR and direct-method phases used in map calculation were weighted by $m_{\mathbf{H}}$ of equation (3).

| | $F_o$-weighted average phase error (°) | Map correlation coefficient |
|---|---|---|
| Unresolved SIR phases | 66.5 | 0.4514 |
| Cluster 1 | | |
| Best set | 54.5 | 0.5144 |
| Average set | 54.5 | 0.5554 |
| Cluster 2 | | |
| Best set | 54.1 | 0.5358 |
| Average set | 54.1 | 0.5609 |
| *DM* based on the average set of cluster 2 | 51.3 | 0.6013 |

showed that they are enantiomorphs of each other. $F$(obs)-weighted phase errors were calculated as

$$\mathrm{ERR} = \sum_{\mathbf{H}} F_{\mathbf{H}}(\mathrm{obs})[|\varphi_{\mathbf{H}}(\mathrm{obs}) - \varphi_{\mathbf{H}}(\mathrm{cal})|]/\sum_{\mathbf{H}} F_{\mathbf{H}}(\mathrm{obs}). \quad (10)$$

Phase errors and map correlation coefficients of the best set and the average set within the biggest two clusters are listed in columns 3–6 of Table 1. The density-modification program *DM* in the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994) was used to improve the direct-method phases. It was observed that SIR phases could not be improved by the *DM* approach, as the phase ambiguity was not resolved. A solvent content of 25% was assumed, and the results are listed in column 7 of Table 1. A typical portion of the electron-density map calculated using SIR phases and using direct method plus *DM* phases are shown in Fig. 1. The quality of the electron-density map calculated using direct method plus *DM* phases should enable one to carry out straightforward model building and structure determination.

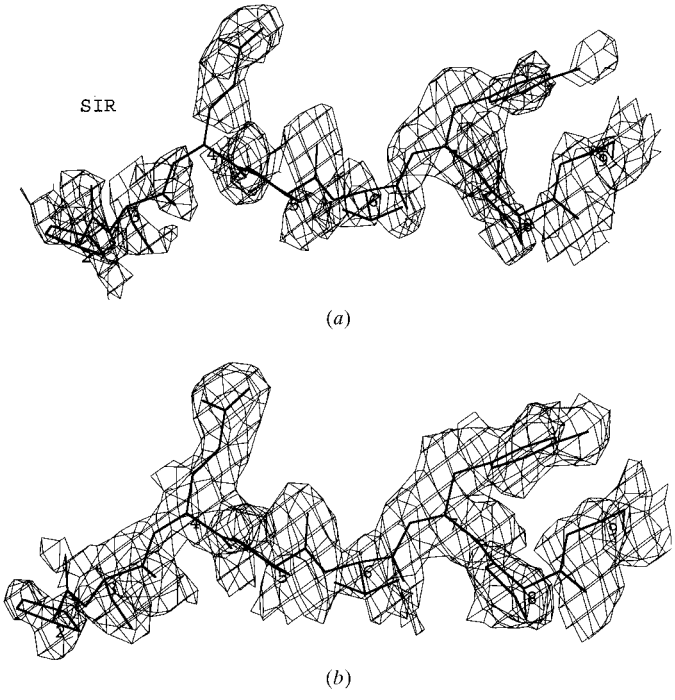


**Figure 1**
A portion of the electron-density map calculated using (*a*) SIR phases and (*b*) direct method plus density modification (*DM*) phases, with the known structure model (Blundell *et al.*, 1981) superimposed.

## 5. Concluding remarks

As can be seen in Table 1, the $F_o$-weighted average phase error for the original SIR phases (Wang, 1981) is 67°, and dropped by 12° to 55° after the direct-methods procedure as described in the previous sections. This means that the SIR phase ambiguity has been effectively broken by the direct method. Unlike the use of figures of merit, cluster analysis could not predict which individual phase set is the best. However, phases as good as those of the 'best set' can be derived by averaging over the 'best cluster' (see corresponding values in Table 1). Finally, while density-modification techniques alone are not able to break the SIR phase ambiguity when the replacing atoms are in a centrosymmetric arrangement, the combination of direct methods and density modification gives much improved results.

FH-F would like to thank Professor T. Blundell for making available the aPP data. The project is supported by the Chinese Academy of Sciences and the National Natural Sciences Foundation of China.

## References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P. & Wu, C. W. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 4175–4179.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Coulter, C. L. (1965). *J. Mol. Biol.* **12**, 292–295.
Fan, H.-F. (1965). *Acta Phys. Sin.* **21**, 1114–1118. (In Chinese.) *Chinese Phys.* pp. 1429–1435. (In English.)
Fan, H.-F. & Gu, Y.-X. (1985). *Acta Cryst.* A**41**, 280–284.
Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.
Giacovazzo, C., Cascarano, G. & Zheng, C. (1988). *Acta Cryst.* A**44**, 45–51.
Hao, Q., Gu, Y.-X., Zheng, C.-D. & Fan, H.-F. (1996). *OASIS: a Computer Program for Breaking the Phase Ambiguity in OAS or SIR Protein Data.* School of Applied Sciences, De Montfort University, Leicester, England and Institute of Physics, Chinese Academy of Sciences, Beijing, People's Republic of China.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.
Hendrickson, W. A. (1971). *Acta Cryst.* B**27**, 1474–1475.
Karle, J. (1966). *Acta Cryst.* **21**, 273–276.
Klop, E. A., Krabbendam, H. & Kroon, J. (1987). *Acta Cryst.* A**43**, 810–820.
Kyriakidis, C. E., Peschar, R. & Schenk, H. (1993). *Acta Cryst.* A**49**, 557–569.
Lunin, V. Yu., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540–544.
Wang, B. C. (1981). *Acta Cryst.* A**37**, C11.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
Yao, J.-X. & Fan, H.-F. (1985). *Acta Cryst.* A**41**, 284–285.