Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Application of direct methods to protein crystallographic phase extension

Fan Jiang

Institute of Physics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China, and Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, People's Republic of China

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

Received 23 October 2001
Accepted 29 May 2002

An alternative formulation of direct methods for phase refinement using the sign probability originally introduced for breaking the phase ambiguity intrinsic in single isomorphous replacement or single-wavelength anomalous diffraction data has been proposed. This formulation can incorporate known phase information through constrained phase refinement. This formulation was tested by application to the phase-extension problem using experimental data. The results showed that this formulation was valid and promising for designing better methods for phase extension.

1. Introduction

The application of direct methods to protein X-ray crystallography has been the product of many years of study and tremendous progress has been made in recent years (Hauptman, 1997). For *ab initio* phasing, the minimal principle has been formulated and applied with success in the *Shake-and-Bake* algorithm, which has been shown to be able to solve the phase problem automatically for structures containing as many as 1000 atoms when data at atomic resolution are available (Miller *et al.*, 1993; Weeks *et al.*, 1995). Approaches to resolve reflection phase ambiguity involved the combination of direct methods with the single-wavelength anomalous scattering method (OAS or SAS; Fan, 1965; Fan & Gu, 1985; Liu, Harvey *et al.*, 1999) or with multi-wavelength anomalous diffraction (MAD) data (Gu *et al.*, 2001). In both cases, the incorporation of direct methods led to better phases and maps with experimental data to medium resolution. These developments will continue as the structural genomics field matures, as the phase problem is still one of the essential steps in determining protein structures in an automated fashion.

It is well known that direct methods can be applied to phase extension. In this work, a phase-refinement formula with constraints on a known phase set was first derived. This formula was then applied to phase extension and tested on a myoglobin structure with experimental data to 2.0 Å resolution. The resulting phase set is shown to be better than that from the program *DM* (density modification) alone in the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). The proposed phase-refinement formula is based on the sign formula for breaking the phase ambiguity in the OAS case (Fan & Gu, 1985). The physical meaning of the sign formula is reinterpreted so that it could be used as a phase-refinement formula through iteration; that is, phases could be refined through a series of steps. The phase shifts are split into magnitudes and signs, which are treated separately by a phase-shift tangent formula and the sign probability. This procedure will be shown to lead to an efficient algorithm for phase refinement which converges.

2. Materials and methods

2.1. Sign formula for single-wavelength anomalous scattering (OAS or SAS) and single isomorphous replacement method (SIR)

The phase ambiguity in the OAS case is expressed as

$$\varphi_h = \varphi''_{h,A} \pm |\Delta\varphi_h|, \tag{1}$$

where $\varphi''_{h,A}$ and $\Delta\varphi_h$ are known from the OAS data (Kartha, 1975). The sign of $\Delta\varphi_h$ is derived using the formula

$$P_+(\Delta\varphi_h) = 0.5 + 0.5\tanh\Big[\sin|\Delta\varphi_h| \sum_{h'} m_{h'} m_{h-h'} \kappa_{h,h'} \times \sin(\Phi'_3 + \Delta\varphi_{h',\text{best}} + \Delta\varphi_{h-h',\text{best}}) + \chi_h \sin\delta_h\Big], \tag{2}$$

where

$$\Phi'_3 = -\varphi''_{h,A} + \varphi''_{h',A} + \varphi''_{h-h',A} \tag{3}$$

and $\chi\sin\delta_h$ is the Sim weight (Sim, 1959),

$$\Delta\varphi_{h,\text{best}} = \varphi_{h,\text{best}} - \varphi''_{h,A}. \tag{4}$$

The values of $\Delta\varphi_{h,\text{best}}$ and m_h are calculated with the initial P_+ set to 0.5,

$$\tan(\Delta\varphi_{h,\text{best}}) = 2(P_+ - \tfrac{1}{2})\sin|\Delta\varphi_h| / \cos\Delta\varphi_h, \tag{5}$$

$$m_h = \exp(-\sigma_h^2/2)\{[2(P_+ - \tfrac{1}{2})^2 + \tfrac{1}{2}](1 - \cos 2\Delta\varphi_h) + \cos 2\Delta\varphi_h\}^{1/2}, \tag{6}$$

where σ_h is related to the experimental error and can be calculated from the mean square of the 'lack-of-closure error' (Blow & Crick, 1959). For the theory behind and the practical application of (1) to (6), the reader is referred to Fan & Gu (1985), Fan *et al.* (1990) and Hao (2000).

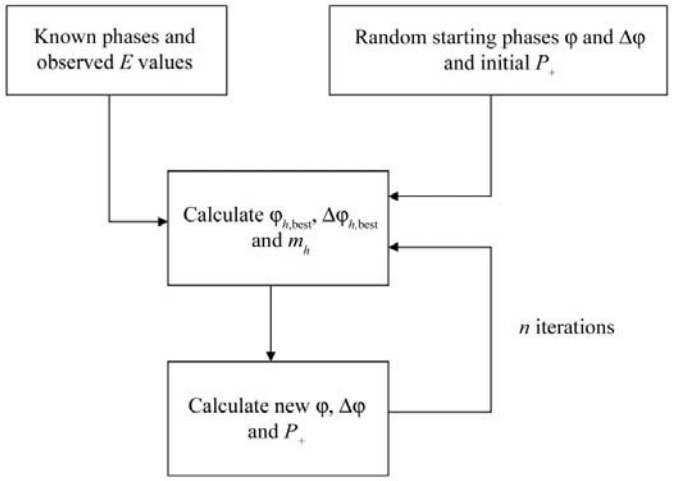

Figure 1
Flow chart of the constrained phase refinement

2.2. Reinterpretation of the physical meaning of the sign formula and its modification

Suppose the true phase is $\varphi_{h,\text{true}}$; for a given trial phase $\varphi_{h,\text{trial}}$ and phase step size $\Delta\varphi_{h,\text{trial}}$, the improved phase $\varphi_{h,\text{imp}}$ is equal to $\varphi_{h,\text{trial}} \pm |\Delta\varphi_{h,\text{trial}}|$. For the OAS case, $\varphi_{h,\text{trial}}$ and $\Delta\varphi_{h,\text{trial}}$ can be calculated from the experimental data and hence are known. When they are not known, random phases are used as the starting values. Direct methods state that the phase information is contained in the diffraction intensity. Therefore, if $\varphi_{h,\text{trial}}$ and $\Delta\varphi_{h,\text{trial}}$ are substituted for $\varphi''_{h,A}$ and $\Delta\varphi_h$ in OAS, respectively, (2) could be used to calculate P_+ and this probability should contain the phase information from direct methods, because $\varphi_{h,\text{imp}}$ depends on P_+ through (4) and (5). It is expected that $\varphi_{h,\text{imp}}$ will gradually approach $\varphi_{h,\text{true}}$. To improve the value of $\Delta\varphi_{h,\text{trial}}$ for each iteration, the following formula derived in Fan & Gu (1985) is used,

$$\tan(\Delta\varphi_{h,\text{trial}}) = \frac{\sum_{h'} m_{h'} m_{h-h'} \kappa_{h,h'} \sin(\Phi'_3 + \Delta\varphi_{h',\text{best}} + \Delta\varphi_{h-h',\text{best}}) + \chi_h \sin\delta_h}{\sum_{h'} m_{h'} m_{h-h'} \kappa_{h,h'} \cos(\Phi'_3 + \Delta\varphi_{h',\text{best}} + \Delta\varphi_{h-h',\text{best}}) + \chi_h \cos\delta_h}. \tag{7}$$

The absolute value of the resulting $\Delta\varphi_{h,\text{trial}}$ is substituted back into (2) and then the resulting value of P_+ into (6) and (5). The value of $\varphi_{h,\text{trial}}$ is then updated with the calculated value of $\Delta\varphi_{h,\text{best}}$. This completes a cycle of iteration, as shown in the flow chart in Fig. 1. In the original derivation of (7) (Fan & Gu, 1985), the Sim-weight distribution was used to introduce the known phase information from the anomalous scatters. In the current reinterpretation of (7), the real physical meaning of the Sim weight is not used. Instead, known phase information $\varphi_{h,\text{known}}$, with a figure of merit m_h, is first approximated by a Gaussian distribution and this Gaussian distribution is then approximated, mathematically and numerically, by a Sim-weight distribution using a Taylor expansion at $\delta_h = 0$, so that (7) could still be used. Therefore, χ_h is approximated by (11) (see below) and δ_h should be $\varphi_{h,\text{known}} - \varphi_{h,\text{trial}}$. The Sim-weight term is only present and added to the triplet term when there is known phase information for the corresponding index h. Therefore, when there is no known phase information, the corresponding Sim-weight term is missing and no phase constraint is applied.

When there is no heavy-atom or anomalous scatter in the crystal structure, the 'lack-of-closure error' σ_h in (6) cannot be calculated as before. However, the phase information from the diffraction intensity also contains error, which is related to the half-height width of the Cochran distribution. To estimate this error, an analogy between the minimal principle (Hauptman, 1991) and the least-squares minimization is made, which is evident by comparing the following two equations.

$$R_{\min} = \sum_{h,h'} \kappa_{h,h'} \left[\cos\Phi_3 - \frac{I_1(\kappa_{h,h'})}{I_0(\kappa_{h,h'})}\right]^2, \tag{8}$$

$$R = \sum_i \frac{1}{\sigma_i^2}(x_i - \bar{x})^2. \quad (9)$$

Therefore, σ_h can be estimated by the following formula,

$$\sigma_h^2 = \frac{1}{\sum_{h'} \kappa_{h,h'}}. \quad (10)$$

For a known phase, $\varphi_{h,\mathrm{known}}$, the Sim-weight-like distribution is used to approximate the phase-probability distribution and χ_h can be calculated by approximating the distribution with the Gaussian distribution whose half-height width is related to the figure-of-merit of the known phase. This leads to the next equation,

$$\chi_h = -1/[2ln(m_h)], \quad (11)$$

where m_h is the figure of merit. There is a singularity in χ_h when m_h approaches zero, which is avoided by setting m_h equal to 1 minus a tiny number. This tiny number will affect the relative weight between the Sim-weight term and the triplet term. This tiny number is chosen to be 10^{-4} and, in the meantime, a scale factor is also introduced to multiply the resulting χ_h in order to modulate the relative balance between the two terms. In the test presented below, the scale factor was 1.0.

2.3. Application to phase extension

When a set of known phases is available with a figure of merit for each phase, this information can be incorporated into phase refinement through the term $\chi \sin \delta_h$ in (2) and (7) using (11). This is especially useful when the known set consists of low-resolution reflections, so that high-resolution phases will become more accurate from constrained phase refinement than that when the known set is not available, which should be random phases, assuming the density-modification procedure has not been applied.

The proposed method was tested with a myoglobin structure with experimental data available from the PDB (Bernstein *et al.*, 1977), PDB code 1dti. It belonged to space group $P2_12_12_1$, with unit-cell parameters $a = 49.163$, $b = 40.002$, $c = 80.011$ Å and one molecule in the asymmetric unit. A set of phases to 4 Å was first generated from the crystal structure using *CCP*4/*SFALL* (Collaborative Computational Project, Number 4, 1994), assigning a figure of merit to 1.0 for all reflections. This set was used as the known set. The experimental data of 1dti used is up to 2.0 Å, the normalized structure amplitudes E of which were calculated with *CCP*4/*ECALC*. The constrained phase refinement was run for 20 iterations for a given random seed, which was used to generate the random starting phases for all the reflections to 2.0 Å. A series of runs were tried and the resulting phase sets were averaged by summation using

$$\varphi_{h,\mathrm{average}} = \tan^{-1}\left[\frac{\sum_{j=1}^{n}(m_h \sin\varphi_{h,\mathrm{best}})_j}{\sum_{j=1}^{n}(m_h \cos\varphi_{h,\mathrm{best}})_j}\right],$$
$$(m_h)_{\mathrm{average}} = \frac{\left\{\left[\sum_{j=1}^{n}(m_h \sin\varphi_{h,\mathrm{best}})_j\right]^2 + \left[\sum_{j=1}^{n}(m_h \cos\varphi_{h,\mathrm{best}})_j\right]^2\right\}^{-1/2}}{n}, \quad (12)$$

where n is the number of phase sets included in the average. The averaged phase set was subject to a standard run of *CCP*4/*DM* and the results were compared with the set from *CCP*4/*DM* alone by checking the map correlation coefficient, phase error and electron-density map connectivity.

The phase sets for average were chosen based on clustering analysis, shown to be effective previously (Liu, Gu *et al.*, 1999). The phase distances between all pairs of phase sets were first calculated and then subject to clustering analysis with the complete linkage algorithm as implemented in the program *OC* (Barton, 1993). The distance between two phase sets is defined as

$$\Delta\varphi = \frac{\sum_i^n m_{1,i}m_{2,i}|\varphi_{1,i} - \varphi_{2,i}|}{\sum_i^n m_{1,i}m_{2,i}}, \quad (13)$$

where m is the figure of merit, subscripts 1 and 2 correspond to phase sets 1 and 2, and subscript i corresponds to the reflection number i.

3. Results and discussion

The test was performed by extending a set of calculated phases from 4.0 to 2.0 Å using the experimental data. There were 1496 known phases and 10 898 observed reflections included. The number of the largest Σ_2 relations included in the calculation was 1 346 541. There were 20 iterations for each run, which took about 10 min of CPU time on an 800 MHz Pentium III. 2000 runs were performed for the present test.

The improvement was certain when the electron-density map connectivity was inspected at 3.5 Å for *DM* alone and the constrained phase refinement followed by *DM*. In the former there were three small breaks, whereas in the latter there were two small breaks. The difference at the density map of residue Ala60 was the largest break in the *DM*-only map and showed that the constrained phase refinement improved the electron-density map (see Fig. 2). The map correlation coefficients and the averaged phase errors calculated at 3.5 Å, however, did not seem very sensitive to the local improvement in the map. The overall (including both the main-chain and side-chain atoms) map correlation coefficient improved only 0.3%, from 0.7902 for *DM* alone to 0.7928 for direct methods plus *DM*, while the averaged phase error improved 1%, from 20.79 for *DM* alone to 20.57 for direct methods plus *DM*. Owing to the fact that *DM* is extremely effective in extending phases that are very accurate at low resolution, particularly for the main-

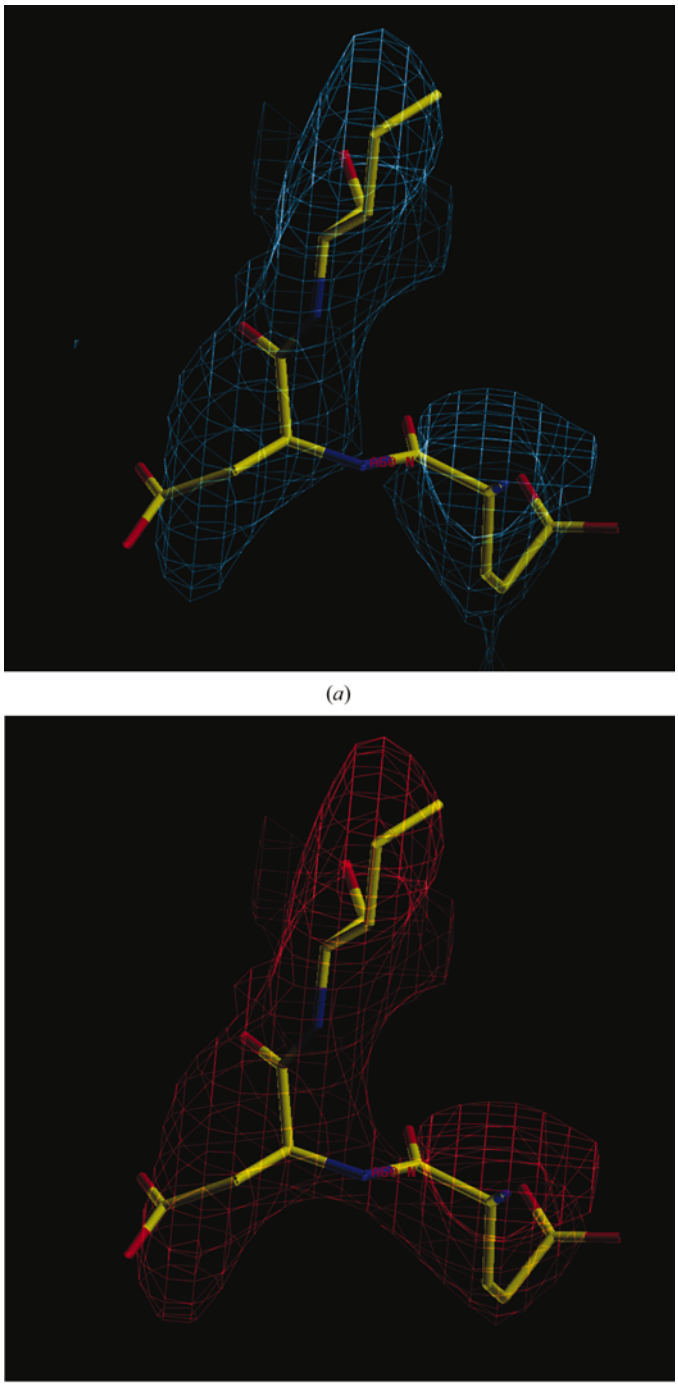

Figure 2
Comparison of electron-density connectivity between *DM* alone and the constrained phase refinement plus *DM*. The map for *DM* alone is in blue (top) and that for the constrained phase refinement in orange (bottom). Both maps are contoured at 1.2σ. The largest break found in the DM map is at the amide N atom of residue Ala60, which is labelled.

chain atoms, it is very difficult to outperform *DM*. Since the calculated phases from the model were used as the seed for phase extension, the improvement shown here is not as significant as one would expect. However, this new method does show some promising aspects for future applications.

Finally, it worth pointing out that if the quality of a phase set could be estimated accurately using an indicator, the good sets of phases could be selected for averaging by clustering analysis (Liu, Gu *et al.*, 1999). In the present test, the phase sets were clustered using the complete linkage algorithm as implemented in the program *OC* (Barton, 1993). A cluster consisting of 11 phase sets was chosen which gave the results discussed above and in Fig. 2. It is interesting to note that the single linkage algorithm for clustering did not give satisfactory results.

Introducing phase information from direct methods is certainly useful and should be attempted whenever possible, but it should be performed in combination with the density-modification procedure, a well established method in protein crystallography. Direct methods extracts phase information from the diffraction amplitudes, so it should be applied first and followed by *DM*, so that the phases input to *DM* contain more phase information. Usually, *DM* will generate the electron-density map for interpretation and tracing. However, further cycling of applying direct methods and *DM* may be attempted in order to determine if further improvement of the electron-density map is possible. During cycling, the convergence of the electron-density maps between cycles could be used as a criterion for the convergence of this new procedure.

In summary, based on the comparison of electron-density maps along the whole molecule, the proposed method performed better than *DM* alone in terms of map connectivity in particular parts of the maps that were compared. The global map correlation and phase-error indicators proved insensitive to these local benefits.

I would like to thank Professor Hai-Fu Fan for introducing me to the sign-probability formula for breaking the phase ambiguity and many insightful suggestions during this work. Project 30170198 was supported by NSFC.

References

Barton, G. J. (1993). *OC: a Cluster Analysis Program.* University of Dundee, Scotland.
Bernstein, F. C., Koetzle, T. F., Williams, J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Fan, H. F. (1965). *Acta Phys. Sin.* **21**, 1114–1118.
Fan, H. F. & Gu, Y. X. (1985). *Acta Cryst.* A**41**, 280–284.
Fan, H. F., Hao, Q., Gu, Y. X., Qian, J. Z., Zheng, C. D. & Ke, H. (1990). *Acta Cryst.* A**46**, 935–939.
Gu, X. Y., Liu, Y. D., Hao, Q., Ealick, S. E. & Fan, H. F. (2001). *Acta Cryst.* D**57**, 250–253.
Hao, Q. (2000). *J. Synchrotron Rad.* **7**, 148–151.
Hauptman, H. A. (1991). *Crystallographic Computing 5: From Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 324–332. Oxford: IUCr/Oxford University Press.
Hauptman, H. A. (1997). *Curr. Opin. Struct. Biol.* **7**, 672–680.
Kartha, G. (1975). *Anomalous Scattering*, edited by S. Ramaseshan & S. C. Abrahams, p. 369. Copenhagen: Munksgaard.
Liu, Y. D., Gu, Y. X., Zheng, C. D., Hao, Q. & Fan, H. F. (1999). *Acta Cryst.* D**55**, 846–848.

Liu, Y. D., Harvey, I., Gu, Y. X., Zheng, C. D., He, Y. Z., Fan, H. F., Hasnain, S. S. & Hao, Q. (1999). *Acta Cryst.* D**55**, 1620–1622.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.