Direct-method phasing of anomalous diffraction from proteins

Y. X. Gu[I], F. Jiang[I], B. D. Sha[II] and H. F. Fan*[,I]

[I] Chinese Academy of Sciences, Institute of Physics, Beijing 100080, P.R. China
[II] University of Alabama at Birmingham, Center for Biophysical Sciences and Engineering, AL 35294-0005, USA

Received ▮▮▮; accepted August 29, 2002

Abstract. Direct methods have been successfully used to break the phase ambiguity intrinsic in single-wavelength anomalous diffraction (SAD) as well as to improve phases from the conventional phasing of multi-wavelength anomalous diffraction (MAD) of proteins.

In dealing with anomalous diffraction data from proteins, the probability distribution of three-phase structure invariants provides additional information, which can be used to resolve the phase ambiguity intrinsic in single-wavelength anomalous diffraction (SAD) enabling solution of the protein structure. Based on this, direct methods have also been used in phasing multi-wavelength anomalous diffraction (MAD) data leading to better result than that from the conventional MAD phasing.

Introduction

Long before synchrotron radiation sources being commonly available in crystallography, the use of anomalous diffraction in single-crystal structure analysis was initiated by Ramachandran and Raman (1956) with single-wavelength anomalous diffraction (SAD) and by Raman (1959) with multi-wavelength (2-wavelength) anomalous diffraction (MAD). With MAD, it is some times difficult to choose suitable wavelengths in the experiment or the sample is not stable for long exposure to X-rays. In contrast, a SAD experiment does not have critical requirement on choosing wavelength and it needs much shorter exposure time. However SAD gives rise to the phase ambiguity obstructing solution of the structure. Ramachandran and Raman (1956) proposed the use of heavy-atom method to break the phase ambiguity. Fan (1965a, b) proposed the use of direct methods instead. Karle (1966) reported a similar method. Hazell (1970), Sikka (1973) Heinerman, Krabbendam, Kroon and Spek (1978) tried to use direct methods in some ways for the same purpose. Extensive study on direct-method phasing of SAD data has been carried out since the 1980's. Hauptman (1982) and Giacovazzo (1983) integrated direct methods with SAD data to solve the phase problem for proteins. The method of Fan (1965a, b) was generalised and incorporated with the treatment of lack-of-closure error used in protein crystallography and the Sim distribution from the partial-structure of anomalous scatterers (Fan *et al.*, 1984a, b; Fan and Gu, 1985). This method has been successfully tested with a number of known proteins (Fan *et al.*, 1990; Sha *et al.*, 1995; Zheng *et al.*, 1996). It has been also applied to solve an originally unknown protein, rusticyanine, with a molecular weight of 16.8 kDa at 2.1 Å resolution (Harvey *et al.*, 1998; Liu *et al.*, 1999). A Program *OASIS* (Hao *et al.*, 2000) based on the procedure of Fan *et al.* (1990) is now available in the latest version of the CCP4 suite (Collaborative Computational Project, Number 4, 1994) for phasing SAD and SIR (single isomorphous replacement) protein data. In a different context, the procedure of Fan *et al.* (1990) has been combined with the conventional MAD phasing leading to the procedure of Direct-method aided MAD phasing (DMAD). Preliminary test has been reported by Gu *et al.* (2001). Further test on a known protein showed that even with 2-wavelength data, the DMAD procedure could lead to an evidently better result than that from a 4-wavelength conventional MAD phasing.

Direct-method phasing of SAD data

The phase ambiguity

In the case of anomalous diffraction, we have

$$F^+ = F^o + F' + F'' \tag{1}$$

$$F^{-*} = F^o + F' - F'' \,, \tag{2}$$

where F^+ and F^- are structure factors of a Friedel pair, F^{-*} is the complex conjugate of F^-, F^o is the structure factor with all atoms in the unit cell scattering normally, F' is the real-part correction for the anomalous scatterers and F'' the imaginary-part correction for the same set of anomalous scatterers. Subtracting (2) from (1) we obtain

$$F^+ - F^{-*} = 2F'' \,. \tag{3}$$

The magnitudes of F^+ and F^- can be measured from experiment. They can then be used to derive the vector F''. Equation (3) implies a triangle shown with solid lines in Fig. 1. However, since we do not know the direction of F^+ and F^-, we have an alternative way to draw the triangle as shown with dashed lines in Fig. 1. This leads to

* Correspondence author (e-mail: fan@aphy.iphy.ac.cn)

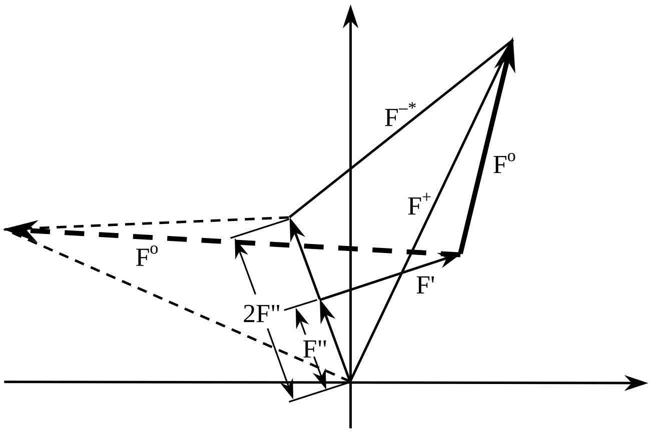

Fig. 1. Two possible solutions for the structure factor F^o resulting from single-wavelength anomalous diffraction. The true solution is shown by a thick solid vector, while the false solution is denoted by a dashed vector.

two possible solutions for the structure factor F^o, indicated by a thick solid vector for the supposed true one and by a dashed vector for the false one.

Define a median structure factor

$$F = F^o + F' . \tag{4}$$

The magnitude of which is given approximately by

$$|F| \simeq (|F^+| + |F^-|)/2 . \tag{5}$$

Corresponding to the two possible solutions of F^o, there are two possible phases for F (see Fig. 2), i.e.

$$\varphi = \varphi'' \pm |\Delta\varphi| , \tag{6}$$

where $|\Delta\varphi|$ can be calculated exactly by solving the triangle in Fig. 2, giving

$$|\Delta\varphi| = \left| \cos^{-1}\left(\frac{|F^+|^2 - |F^-|^2}{4|F||F''|} \right) \right| \tag{7}$$

or approximately as

$$|\Delta\varphi| = \left| \cos^{-1}\left(\frac{|F^+| - |F^-|}{2|F''|} \right) \right| . \tag{8}$$

Equation (6) defines the phase ambiguity arising from single-wavelength anomalous diffraction.

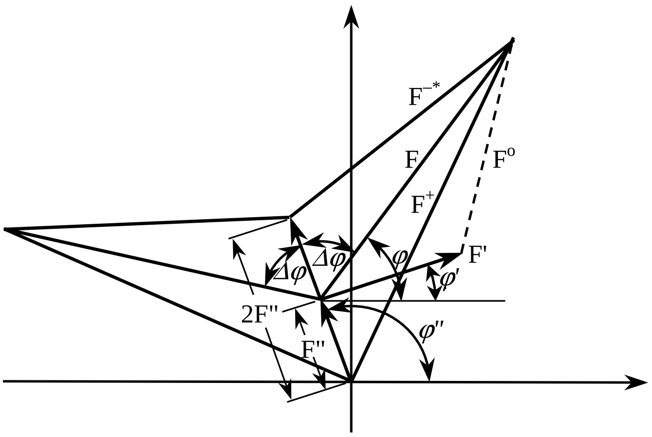

Fig. 2. Phase ambiguity arising from single-wavelength anomalous diffraction.

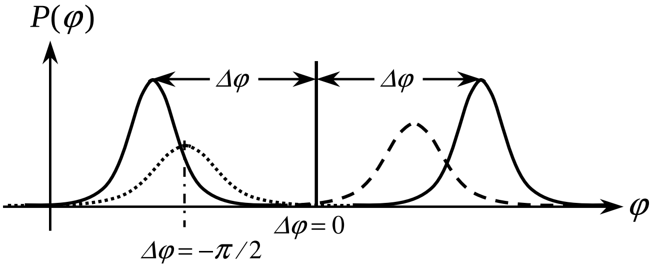

Fig. 3. Bimodal distribution of the phase doublet from SAD (solid line), Sim's distribution (dotted line) and the Cochran distribution (dashed line).

Strategy of breaking the phase ambiguity

Assuming a Gaussian distribution and notice that (see Blundell and Johnson, 1976)

$$\Delta F = F^+ - F^- \simeq 2|F''| \cos \Delta\varphi , \tag{8}$$

the bimodal phase distribution of a particular reflection with reciprocal vector ***h*** in the SAD case can be expressed as

$$P_{\text{anom}}(\varphi_{\mathbf{h}}) = N \exp\left[-\sigma(\Delta F - 2|F''| \cos \Delta\varphi_{\mathbf{h}})^2\right] , \tag{9}$$

where N is the normalising coefficient and σ is related to the variance. The distribution has two identical peaks symmetrically related to the vertical line at $\Delta\varphi = 0$ (see the solid curve in Fig. 3). In order to break the phase ambiguity, we should have some way to modify the distribution so as the two peaks can be different to each other. Since the substructure of the anomalous scatterers is assumed to be known, it can be used to calculate the Sim distribution (Sim, 1959)

$$P_{\text{Sim}}(\varphi_{\mathbf{h}}) = N' \exp\left(-x \sin \Delta\varphi_{\mathbf{h}}\right) , \tag{10}$$

where N' is the normalising coefficient and x is related to the magnitude of the structure factor of the whole unit cell and that of the substructure of anomalous scatterers. Evidently, at least in theory, the phase ambiguity can be resolved by multiplying formulas (9) and (10). This is actually the starting point of some practical phasing procedures for dealing with SAD data. However, since the Sim distribution is peaked at $\Delta\varphi = -\pi/2$, it always enhances the left peak of the bimodal distribution and could never increase the relative height of the right (see Fig. 3). On the other hand, according to the Cochran distribution (Cochran, 1955), we have

$$P_{\text{Cochran}}(\varphi_{\mathbf{h}}) = N'' \exp\left[\sum_{\mathbf{h}} \varkappa \cos\left(\varphi_{\mathbf{h}} - \varphi_{\mathbf{h}'} - \varphi_{\mathbf{h}-\mathbf{h}'}\right)\right] , \tag{11}$$

where N'' is the normalising coefficient and $\varkappa$ is related to the magnitude of three-phase structure invariants involving ***h***, ***h***$'$ and ***h*** $-$ ***h***$'$. Unlike the Sim distribution, Cochran's distribution can have a maximum anywhere from $\varphi = 0$ to $\varphi = 2\pi$. Hence, as a starting point for breaking the phase ambiguity of SAD data, the product of formulas (9), (10) and (11) will be definitely superior to that involving only formulas (9) and (10).

Direct-method phasing procedure

Based on the above consideration, direct-method formulas have been derived for resolving the phase ambiguity intrinsic in SAD data (see Fan and Gu, 1985 and references therein). The phasing procedure is summarised as follows.

(i) Phase doublets from SAD data are expressed as

$$\varphi_{\mathbf{h}} = \varphi_{\mathbf{h}} \pm |\Delta\varphi_{\mathbf{h}}| \,. \tag{12}$$

(ii) The probability of $\Delta\varphi_{\mathbf{h}}$ to be positive is predicted by the following formula.

$$\begin{aligned} &P_{+}(\Delta\varphi_{\mathbf{h}}) \\ &= \frac{1}{2} + \frac{1}{2}\ \tanh\left\{\sin\left(|\Delta\varphi_{\mathbf{h}}|\right)\left[\sum_{H'} m_{\mathbf{h}'} m_{\mathbf{h}-\mathbf{h}'} \varkappa_{\mathbf{h},\mathbf{h}'}\right.\right. \\ &\left.\left.\times \sin\left(\Phi_3' + \Delta\varphi_{\mathbf{h}',best} + \Delta\varphi_{\mathbf{h}-\mathbf{h}',best}\right) + \chi \sin \delta_{\mathbf{h}}\right]\right\} \end{aligned} \tag{13}$$

(iii) The best phase and figure of merit of each reflection are than calculated as respectively

$$\tan\left(\Delta\varphi_{\mathbf{h}\,best}\right) = 2\left(P_{+} - \frac{1}{2}\right) \sin|\Delta\varphi_{\mathbf{h}}| \Big/ \cos \Delta\varphi_{\mathbf{h}} \tag{14}$$

and

$$\begin{aligned} m_{\mathbf{h}} = \exp\left(-\sigma_{\mathbf{h}}^2/2\right) &\left\{\left[2\left(P_{+} - \frac{1}{2}\right)^2 + \frac{1}{2}\right]\right. \\ &\left.\times\left(1 - \cos 2\Delta\varphi_{\mathbf{h}}\right) + \cos 2\Delta\varphi_{\mathbf{h}}\right\}^{1/2} . \end{aligned} \tag{15}$$

(iv) A Fourier map is than calculated and improved by a density-modification procedure.

The procedure has been tested with a number of known proteins (Fan *et al.*, 1990; Sha *et al.*, 1995; Zheng *et al.*, 1996) and used to solve an originally unknown protein (Harvey *et al.*, 1998; Liu *et al.*, 1999). In all cases, a comparison was made between the procedures with and without direct methods. The result of the procedure with direct methods is always the better.

Direct-method aided MAD phasing

A set of MAD data consists of several sets of SAD data at different wavelengths. The conventional phasing of MAD data is to combine the bimodal phase distribution of the corresponding SAD data sets to give a unique phase indication for individual reflections. Since the anomalous scattering effect is weak. the resultant phase indication is also weak. Typically, for a protein of moderate size with diffraction at 2 ~ 3 Å resolution, about one third of the total reflections may have MAD phases with a figure of merit less than 0.3. This explains why in some cases MAD phases are not good enough to initiate a phase-improvement procedure such as solvent flattening. On the other hand, direct methods provide independent phase information without requiring additional experimental data. It can be use to improve phases from conventional MAD phasing. The DMAD (Direct-method aided MAD phasing) procedure is designed for this purpose. The technique is based on both the conventional MAD phasing and the direct-method treatment of SAD data.

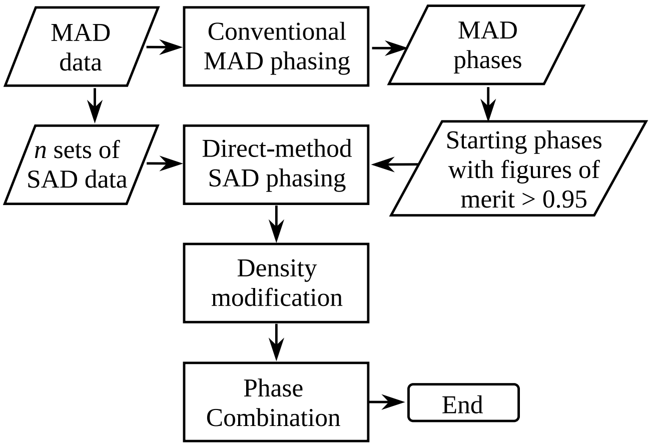

Fig. 4. The flow chart of DMAD procedure.

Phasing strategy

The flow chart of the DMAD procedure is shown in Fig. 4. The main points are:

(i) The conventional MAD phasing is first applied to a set of MAD data.
(ii) MAD phases with figures of merit larger than a certain limit, say 0.95 are used as starting phases in the direct method phasing.
(iii) The MAD data are divided into n sets of SAD data
(iv) Direct methods are used to break the phase ambiguity of each set of SAD data based on the starting phases from the conventional MAD phasing.
(v) Combine results from the n sets of SAD data.
(vi) Combine the combined SAD phases with the conventional MAD phases.

The phase combination is performed according to the following formulas.

$$\varphi_{\text{Combined}} = \tan^{-1}\left[\frac{\sum\limits_{j=1}^{n}\left(m_{\mathbf{h}} \sin \varphi_{\text{best}}\right)_j}{\sum\limits_{j=1}^{n}\left(m_{\mathbf{h}} \cos \varphi_{\text{best}}\right)_j}\right] , \tag{16}$$

$$\begin{aligned} &(m_{\mathbf{h}})_{\text{Combined}} \\ &= \frac{\left\{\left[\sum\limits_{j=1}^{n}\left(m_{\mathbf{h}} \sin \varphi_{\text{best}}\right)_j\right]^2 + \left[\sum\limits_{j=1}^{n}\left(m_{\mathbf{h}} \cos \varphi_{\text{best}}\right)_j\right]^2\right\}^{1/2}}{n} , \end{aligned} \tag{17}$$

where n is the number of phase sets involved in the combination. Such a combination can be regarded as a reciprocal-space equivalent of calculating a sum function of Fourier maps corresponding to n sets of phases.

Data and test results

Preliminary test of the DMAD procedure has been reported by Gu et al. (2001). A further test with another

Table 1. Summary of the test data from the protein, yeast Hsp40 protein Sis1.

Space group	$P4_12_12$
Unit-cell	$a = 73.63$, $c = 80.76$ Å
Independent non-H atoms	1380
Number of independent Se sites	1
Wavelength (Å)	1.0688, 0.9798, 0.9794, 0.9253
Resolution	30 ~ 3.0 Å
Unique reflections	4590

known protein, the yeast Hsp40 protein Sis1 (Sha *et al.*, 2000; see Table 1 for a summary), is described below.

Reflections of the 4-wavelength data were first treated by the conventional MAD procedure followed by density modification with the program 'dm' in the CCP4 suite (Collaborative Computational Project, Number 4, 1994). Reflections from the subset of 2-wavelength data ($\lambda = 0.9798$Å and 0.9794 Å) were treated separately in the same way. The two sets of results were sorted in descending order of F(obs) and cumulated into 9 groups as listed in Table 2. It is seen that the result of 4-wavelength data is obviously better than that of 2-wavelength data.

The above MAD phases of 4-wavelength data and 2-wavelength data were improved separately by the DMAD procedure and density modification. F(obs)-weighted phase errors are listed in Table 3 in comparison with results from conventional MAD phasing of the 4-wavelength data.

Table 2. Phase errors resulting from conventional MAD phasing followed by density modification, comparison of results from 4-wavelength data and from 2-wavelength data.

Number of reflections	F(obs)-weighted averaged phase errors	
	4-wavelength data	2-wavelength data
500	19.36	20.90
1000	23.81	25.03
1500	26.04	28.01
2000	28.13	30.14
2500	30.20	32.73
3000	32.35	34.93
3500	34.11	36.88
4000	35.99	38.71
4500	37.40	40.03

Table 3. Phase errors resulting from the DMAD procedure followed by density modification in comparison with that from conventional MAD phasing and density modification.

Number of reflections	Conventional MAD phasing	Direct-method aided MAD phasing	
	4-wavelength data	4-wavelength data	2-wavelength data
500	19.36	16.02	18.28
1000	23.81	21.16	22.57
1500	26.04	23.47	25.73
2000	28.13	25.43	27.45
2500	30.20	27.20	29.78
3000	32.35	28.99	31.50
3500	34.11	30.69	33.22
4000	35.99	32.35	34.94
4500	37.40	33.98	36.36

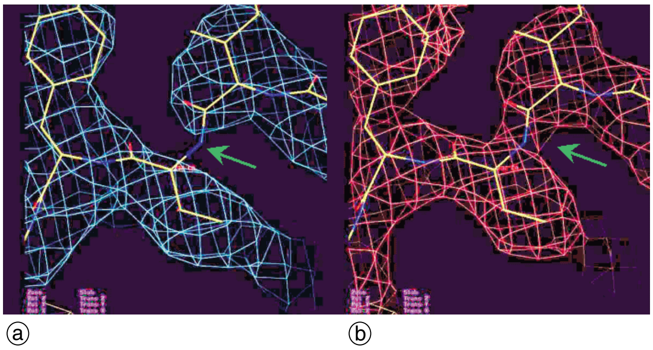

Fig. 5. A portion of the Fourier map of yeast Hsp40 protein Sis1 around the residual 232. (**a**) conventional MAD phaing of 4-wavelength data; (**b**) DMAD phasing of 2-wavelength data.

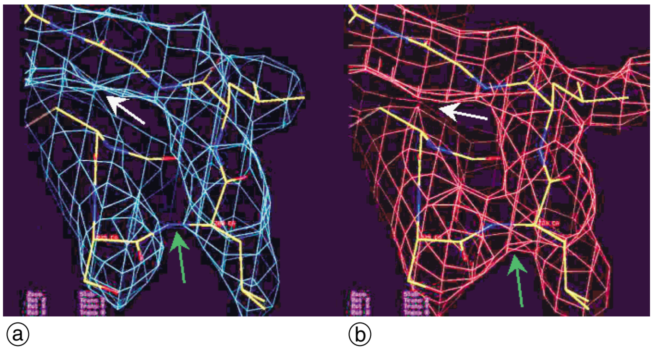

Fig. 6. A portion of the Fourier map of yeast Hsp40 protein Sis1 around the residual 280. (**a**) conventional MAD phasing of 4-wavelength data; (**b**) DMAD phasing of 2-wavelength data.

It is seen that with the 4-wavelength data, the conventional MAD phases have been improved through the DMAD procedure by more than 3 degrees. Moreover, even with the 2-wavelength data, the DMAD procedure could lead to a result better than that from the 4-wavelength conventional MAD phasing. Although the improvement in this case is only about 1 degree, the effect on the corresponding Fourier maps is evident (see Fig. 5 and Fig. 6).

Concluding remarks

Direct methods have been proved powerful for phasing SAD data. They are also useful in dealing with MAD data leading to significantly better results than that from the conventional MAD phasing.

Acknowledgments. Gu, Y. X. and Fan, H. F. should like to thank Prof. M. M. Woolfson, Prof. E. Dodson and Dr. J. X. Yao for their helpful discussion. The project is supported by the Chinese Academy of Sciences, the British Royal Society and by the Ministry of Science and Technology of China, Grant No.: G1999075604.

References

Blundell, T. L.; Johnson, L. N.: Protein Crystallography. Academic Press Inc. London 1976 p. 177.

Cochran, W.: Relations between the phases of structure factors. Acta Cryst. **8** (1955) 473–478.

Collaborative Computational Project, Number 4. Acta Cryst. **D50** (1994) 760–763.

Fan, H. F.: The use of sign relationship in the determination of heavy-atom-containing crystal structures II. Component relationship and its applications. Acta Phys. Sin. **21** (1965a) 1114–1118 (in Chinese).

Fan, H. F.: The use of sign relationship in the determination of heavy-atom-containing crystal structures II. Component relationship and its applications. Chinese Physics (1965b) 1429–1435.

Fan H. F.; Gu, Y. X.: Combining direct methods with isomorphous replacement or anomalous scattering data III. The incorporation of partial structure information. Acta Cryst. **A41** (1985) 280–284.

Fan H. F.; Han F. S.; Qian, J. Z.: Combining direct methods with isomorphous replacement or anomalous scattering data II. The treatment of errors. Acta Cryst. **A40** (1984b) 495–498.

Fan H. F.; Han, F. S.; Qian, J. Z.; Yao, J. X.: Combining direct methods with isomorphous replacement or anomalous scattering data I. Acta Cryst. **A40** (1984a) 489–495.

Fan, H. F.; Hao, Q.; Gu, Y. X.; Qian, J. Z.; Zheng, C. D.; Ke, H.: Combining direct methods with isomorphous replacement or anomalous scattering data VII. Ab-initio phasing of the OAS data from a small protein. Acta Cryst. **A46** (1990) 935–939.

Giacovazzo, C.: The estimation of two-phase and three-phase invariants in *P*1 when anomalous scatterers are present. Acta Cryst. **A39** (1983) 585–592.

Gu, Y. X.; Liu, Y. D.; Hao, Q.; Ealick, S. E.; Fan, H. F. Direct-method-aided phasing of MAD data. Acta Cryst. **D57** (2001) 250–253.

Hao, Q.; Gu, Y. X.; Zheng, C. D.; Fan, H. F.: OASIS – a computer program for breaking the phase ambiguity in one-wavelength anomalous scattering or single isomorphous substitution (replacement) data. J. Appl. Cryst. **33** (2000) 980–981.

Harvey, I.; Hao, Q.; Duke, E. M. H.; Ingledew, W. J.; Hasnain, S. S.: Structure Determination of a 16.8 kDa Copper Protein at 2.1 Å Resolution Using Anomalous Scattering Data with Direct Methods. Acta Cryst. **D54** (1998) 629–635.

Hauptman, H.: On integrating the techniques of direct methods with anomalous dispersion. I. The theoretical basis. Acta Cryst. **A38** (1982) 632–641.

Hazell, A. C.: Structure determination by the combination of anomalous scattering and direct methods. Nature **227** (1970) 269.

Heinerman, J. J. L.; Krabbendam, H.; Kroon, J.; Spek, A. L.: Direct phase determination of triple products from Bijvoet inequalities. II. A probabilistic approach. Acta Cryst. **A34** (1978) 447–450.

Karle, J.: Isomorphous substitution and Formulas for phase determination. Acta Cryst. **21** (1966) 273–276.

Liu, Y. D.; Harvey, I.; Gu, Y. X.; Zheng, C. D.; He, Y. Z.; Fan, H. F.; Hasnain, S. S.; Hao, Q.: Is single-wavelength anomalous scattering sufficient for solving phases? A comparison of different methods for a 2.1 Å structure solution. Acta Cryst. **D55** (1999) 1620–1622.

Ramachandran, J. N.; Raman, S: A new method for the structure analysis of non-centrosymmetric crystals. Curr. Sci. **25** (1956) 348–51.

Raman, S.: Theory of the anomalous dispersion method of determining the structure and absolite configuration of non-centrosymmetric crystals. Proc. Indian Acad. Sci. **A50** (1959) 95–107.

Sha, B. D.; Liu, S. P.; Gu, Y. X.; Fan, H. F.; Ke, H.; Yao, J. X.; Woolfson, M. M.: Direct phasing of one-wavelength anomalous scattering data of the protein core streptavidin. Acta Cryst. **D51** (1995) 342–346.

Sha, B. D.; Lee, S.; Cyr, D. M.: The crystal structure of the peptide-binding fragment from the yeast Hsp40 protein Sis1. Structure **8** (2000) 799–807.

Sikka, S. K.: Use of the tangent formula to resolve the phase ambiguity in the Neutron anomalous-dispersion method. Acta Cryst. **A29** (1973) 211–212.

Sim, G. A.: The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. Acta Cryst. **12** (1959) 813–815.

Zheng, X. F.; Fan, H. F.; Hao, Q.; Dodd, F. E.; Hasnian, S. S.: Direct-method structure determination of the native azurin II protein using one-wavelength anomalous scattering data. Acta Cryst. **D52** (1996) 937–941.