Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 16 October 2003
Accepted 5 January 2004

Statistical approach to the phase problem

Hongliang Xu* and Herbert A. Hauptman

Hauptman–Woodward Medical Research Institute and Department of Structural Biology, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, 73 High Street, Buffalo, NY 14203, USA. Correspondence e-mail: xu@hwi.buffalo.edu

The minimal function and its minimal principle employed in the traditional *Shake-and-Bake* algorithm rely on the probabilistic estimates of the cosines of the structure invariants. In this paper, a novel statistical approach to the phase problem, which utilizes statistical properties of the structure invariants, is proposed. The statistical maximal function and its maximal principle are formulated, and the corresponding statistical *Shake-and-Bake* algorithm and its associated statistical parameter-shift procedure are proposed and tested. The test results show that the statistical approach to the phase problem is a simple, reliable, less computationally intensive and more efficient procedure for phase determination in X-ray crystallography.

1. Introduction

The phase problem of X-ray crystallography may be defined as the problem of determining the phases of the structure factors from measurements of intensities alone. The phase information, which is lost in the diffraction experiment, is in fact recoverable from the measurable intensities. The methods devised to achieve this goal are known as *direct methods*.

1.1. Algebraic background

If **H** is an arbitrary reciprocal-lattice vector, then the structure factor F_H and the normalized structure factor E_H are defined by

$$F_H = |F_H| \exp(i\phi_H) = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \quad (1)$$

$$E_H = |E_H| \exp(i\phi_H) = F_H / \langle |F_H|^2 \rangle^{1/2}, \quad (2)$$

respectively, where N is the number of atoms in the unit cell, f_j and $\mathbf{r}_j$ are the scattering factor and the position vector of the jth atom. Since the position vectors depend on the choice of the origin, the phases are also origin dependent. Certain linear combinations of the phases, the *structure invariants*, are uniquely determined by the structure and are independent of the choice of the origin (Hauptman & Karle, 1953). The most important of these invariants are the triplets

$$\phi_{HK} = \phi_H + \phi_K + \phi_{-H-K}, \quad (3)$$

along with their associated parameters A_{HK}, defined in the equal-atom case by

$$A_{HK} = 2N^{-1/2} |E_H E_K E_{H+K}|. \quad (4)$$

David Sayre, using his algebraic 'squaring' method, which exploits the connection between a crystal structure factor and its squared structure factor, derived the famous Sayre equation (Sayre, 1952),

$$F_H = (V\theta_H)^{-1} \sum_{K=-\infty}^{+\infty} F_K F_{H-K}, \quad (5)$$

where V is the volume of the crystal unit cell and θ_H, in the equal-atom case, is the ratio of two scattering factors between the 'squared' structure and the structure itself. This equation explicitly reveals the algebraic relationship among the triplets *via* measured diffraction intensities.

1.2. Probabilistic background

In the probabilistic approach to the phase problem, the atomic position vectors **r** of the atoms in a crystal are assumed to be random variables, uniformly and independently distributed in the unit cell. The normalized structure factors E, as functions of the atomic position vectors **r**, are themselves random variables. The standard theory of mathematical probability is applied to derive (i) the probability distributions of the structure invariants, and (ii) the conditional probability distributions of the structure invariants, given well defined sets of measured intensities.

The conditional probability distribution, $P(\phi|A_{HK})$, of the triplet ϕ_{HK}, given A_{HK}, is known to be

$$P(\phi|A_{HK}) = [2\pi I_0(A_{HK})]^{-1} \exp(A_{HK} \cos\phi), \quad (6)$$

where I_0 is the modified Bessel function of zero order. From (6), it readily follows that the conditional expected value of ϕ_{HK} is zero. Thus, an estimate of

$$\phi_{HK} = \phi_H + \phi_K + \phi_{-H-K} \approx 0 \quad (7)$$

is valid provided that the values of A_{HK} are large.

1.3. The tangent formula

The first application of the probabilistic approach to the phase problem is the tangent formula (Karle & Hauptman, 1956),

$$\tan(\phi_H) = \frac{\sum_K W_{HK} \sin(\phi_H + \phi_{H-K})}{\sum_K W_{HK} \cos(\phi_H + \phi_{H-K})}, \quad (8)$$

where W_{HK} are appropriate weights. The tangent formula, together with its modified forms, represents the earliest development of the probabilistic approach to the phase problem and demonstrates the power of probabilistic methods on which the direct methods of phase determination are primarily based.

1.4. The minimal principle

As N increases, A_{HK} decreases and the estimate (7) is no longer valid. This limitation has motivated the formulation of a least-squares minimal principle (Debaerdemaeker & Woolfson, 1983) involving the cosine of the structure invariants instead of the structure invariants themselves. The conditional expected value of this cosine is:

$$\langle \cos \phi_{HK} | A_{HK} \rangle = I_1(A_{HK})/I_0(A_{HK}), \quad (9)$$

where I_1/I_0 is the ratio of the modified Bessel functions of the first and zeroth order (Cochran, 1955).

The phase problem is formulated as a problem in constrained global minimization, the constraints arising from identities among the phases that must, of necessity, be satisfied. The commonly used cosine minimal function (DeTitta *et al.*, 1994),

$$R(\phi) = \left(\sum_{H,K} A_{HK} \right)^{-1} \sum_{H,K} A_{HK} \left[\cos(\phi_{HK}) - \frac{I_1(A_{HK})}{I_0(A_{HK})} \right]^2, \quad (10)$$

measures the mean-square difference between the current values of the cosine structure invariants, $\cos(\phi_{HK})$, and their conditional expected values. It is expected that the minimal function (10) reaches its constrained global minimum when all the phases are equal to their true values for any choice of origin and enantiomorph (the minimal principle). However, the complexity of the resulting phase estimation problem is significant because of the existence of multiple local minima owing to the presence of the trigonometric functions. This may severely reduce the radius of convergence of equation (10) and increase the difficulty of reaching the constrained global minimum.

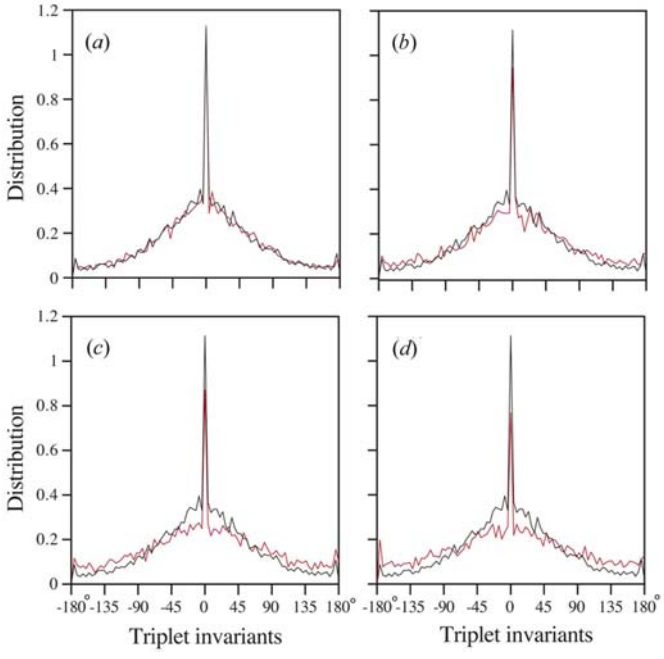

Figure 1
Distribution of triplets for Iled. The black line in (*a*)–(*d*) represents distribution of a phase set with MPE = 0°, the red line represents distribution of a phase set with (*a*) MPE = 10.5°, (*b*) MPE = 32.4°, (*c*) MPE = 56.4° and (*d*) MPE = 75.7°.

2. Statistical approach to the phase problem

In this paper, a novel statistical approach to the phase problem is presented. Unlike the probabilistic approach described in the previous section, the statistical approach utilizes statistical properties of the structure invariants themselves instead of the probabilistic estimates of the cosines of the structure invariants.

In order to study statistical properties of the structure invariants of crystal structures, a number of known structures are chosen as targets. A random-number generator is used to assign uniformly distributed errors to the triplet phases in such a way that sets of triplets are created having the desired mean phase error (MPE), where MPE is the average phase difference between the current phase set and the phase set obtained from the final refined model structure. For each structure, several different sets of phase values are generated with various MPE values. For each generated phase set, the following steps are taken to create a discrete distribution function of the structure invariants:

1. Reflections are sorted in decreasing order of $|E|$s, and a predetermined number of the top reflections are selected to generate a predetermined number of structure invariants having the largest A_{HK} values.

2. The structure invariant interval of $[-180^\circ, 180^\circ]$ is partitioned into a predetermined number of equal subintervals.

3. The value of each structure invariant is calculated and its location (subinterval) is identified.

4. The number of structure invariants in each subinterval is tallied and then divided by the total number of the structure invariants.

5. A discrete graph of the distribution of values of the structure invariants is plotted based on the information obtained from step 4.

It should be pointed out that the discrete distribution is a function of all selected phases. When a particular phase value changes, all triplet values associated with this phase will change, the locations of the triplets will change, as will the shape of the distribution function. Fig. 1 shows distributions of values of triplets for Iled (Pletnev *et al.*, 1980), an 84 atom structure that crystallizes in space group $P2_12_12_1$, as errors are introduced into the values of the phases. The 840 reflections having largest $|E|$ values are chosen to generate 8400 triplets having the largest A values. The triplet interval of $[-180^\circ, 180^\circ]$ is partitioned into 90 equal subintervals. Five

Table 1
Structure data sets used in this investigation.

Structure type	Structure name or ID	Atoms (ASU)	Space group	Resolution (Å)	Reference
Centrosymmetric	STR1	27	$C2/c$	0.83	Camiolo *et al.* (2001)
	STR2	37	$P2_1/c$	0.84	Zhuang *et al.* (2002)
	STR3	46	$P2_1/c$	0.81	Alfonso & Stoeckli-Evans (2001)
	STR4	75	$P\bar{1}$	0.84	Ohba *et al.* (2002)
Non-centrosymmetric	Ph8755	74	$P1$	0.92	Marshall *et al.* (1990)
	Iled	84	$P2_12_12_1$	0.94	Pletnev *et al.* (1980)
	F5	113	$P2_12_12_1$	1.00	Pletnev *et al.* (1992)
	Crambin	327	$P2_1$	0.83	Hendrickson & Teeter (1981)
Se–Met substructure	1JC4	24	$P2_1$	2.00	McCarthy *et al.* (2001)
	1A7A	30	$C222$	2.80	Turner *et al.* (1998)
	1L8A	40	$P2_1$	3.50	Arjunan *et al.* (2002)
	1M32	66	$P2_1$	2.55	Chen *et al.* (2002)

Table 2
The statistical maximal function values for selected (sub)structures The phase sets with different MPE (mean-phase-error) were generated by *SnB*.

STR2		Iled		1A7A	
MPE (°)	$S(\phi)$	MPE (°)	$S(\phi)$	MPE (°)	$S(\phi)$
0.0	0.868	0.0	0.787	0.0	0.768
17.0	0.769	13.2	0.775	10.6	0.772
23.4	0.731	26.1	0.751	18.2	0.763
33.1	0.724	34.0	0.721	29.0	0.752
42.8	0.688	44.3	0.696	32.3	0.727
52.5	0.674	53.5	0.688	48.1	0.693
64.2	0.666	62.8	0.675	54.8	0.684
73.5	0.636	74.7	0.663	68.1	0.660
84.7	0.524	87.5	0.650	78.2	0.656

different sets of phase values (MPE = 0.0, 10.5, 32.4, 56.4 and 75.7°) are used to demonstrate five different distributions: the black lines (*a*)–(*d*) represent the distribution with MPE = 0.0°; the red lines represent distributions with (*a*) MPE = 10.5°, (*b*) MPE = 32.4°, (*c*) MPE = 56.4° and (*d*) MPE = 75.7°.

From Fig. 1, we observe that: (i) a large spike appears in the middle of each of the distributions owing to the large number of triplets having large A values and the property (7); (ii) when the MPE is increased, the height of the spike is correspondingly reduced, the middle portion of the distribution sinks, and the two tails of the distribution rise. Although we present the triplets distribution function for only one particular structure, these observations are the common properties of a dozen test structures consisting of centrosymmetric and non-centrosymmetric structures and Se-atom substructures spanning different sizes, resolutions and space groups. The basic structural information including structure name, number of non-H atoms in the asymmetric unit (ASU), space group, resolution and reference is listed in Table 1.

These statistical properties motivate us to define a maximal function as follows:

$$S(\phi) = \int_{-\pi/2}^{\pi/2} D(\phi_{HK})\, \mathrm{d}\phi_{HK}, \qquad (11)$$

where $D(\phi_{HK})$ is a triplet distribution function on $[-\pi, \pi]$. $S(\phi)$ can also be interpreted as the area bounded by the triplet distribution function $D(\phi_{HK})$, x axis and lines $x = -\pi/2$ and $x = \pi/2$. Let $I = \cup_{j=1}^{M} I_j$ be a partition of $[-\pi/2, \pi/2]$, C_j the number of triplets whose values belong to I_j and C the total number of triplets, then

$$S(\phi) = \int_{-\pi/2}^{\pi/2} D(\phi_{HK})\, \mathrm{d}\phi_{HK} = \sum_{j=1}^{M} \frac{C_j}{C} = \frac{C_I}{C}, \qquad (12)$$

where $C_I = \sum_{j=1}^{M} C_j$ is the number of triplets whose values belong to $[-\pi/2, \pi/2]$.

Table 2 lists statistical maximal function values for selected (sub)structures STR2, Iled and 1A7A. The phase sets with different MPE (mean phase error) were generated using the program *SnB* (Weeks & Miller, 1999*a*) and then used to calculate $S(\phi)$. It is observed that, as expected, there is very good correlation between the values of the statistical maximal function $S(\phi)$ and the mean phase errors, *i.e.*, the larger the value of $S(\phi)$, the smaller the mean phase error. In fact, the correlation coefficients, between the values of the statistical maximal function $S(\phi)$ and the MPE listed in Table 2, are −0.954, −0.981 and −0.974 for STR2, Iled and 1A7A, respectively. In view of Table 2, one naturally anticipates that the statistical maximal function, $S(\phi)$, reaches its constrained global maximum when all phases are equal to their true values for any choice of origin and enantiomorph (the statistical maximal principle). If this hypothesis is true, then the phase problem can be formulated as a problem of constrained global maximization of the statistical maximal function $S(\phi)$. It is worth pointing out that it is essential to find the *constrained* global maximum instead of the (unconstrained) global maximum since the set of phases all zero would maximize $S(\phi)$.

3. Statistical *Shake-and-Bake*

It is one thing to formulate the phase problem as a problem of constrained global optimization, it is quite another to actually find the constrained global maximum. The *Shake-and-Bake* algorithm (Miller *et al.*, 1993; DeTitta *et al.*, 1994; Weeks *et al.*, 1994), the most powerful direct-methods-based algorithm yet devised, shows the way. *Shake-and-Bake*, the first algorithm to find the constrained global minimum of a probabilistically defined minimal function, alternated phase refinement in

Table 3
Values of basic *SnB* parameters.

Structure	Phases	Triplets	Peaks	Cycles
STR1	400	4000	27	13
STR2	370	3700	37	18
STR3	460	4600	46	23
STR4	750	7500	75	37
Ph8755	740	7400	74	37
Iled	840	8400	84	42
F5	1130	11300	90	113
Crambin	3270	32700	100	300
1JC4	840	8400	28	48
1A7A	900	9000	30	60
1L8A	1260	12600	42	84
1M32	1980	19800	66	132

Table 4
Success rates obtained from traditional and statistical *Shake-and-Bake*.

	Success rate (%)	
Structure	Traditional *SnB*	Statistical *SnB*
STR1	1.0	1.8
STR2	4.9	5.9
STR3	1.7	2.5
STR4	1.4	0.8
Ph8755	54.0	60.3
Iled	4.3	5.0
F5	1.3	2.9
Crambin	3.2	2.8
1JC4	27.5	32.1
1A7A	3.9	7.8
1L8A	2.5	4.4
1M32	4.1	6.0

reciprocal space with density modification in real space to impose constraints through a physically meaningful interpretation of the electron-density function. Specifically, the phase-refinement portion of the *Shake-and-Bake* cycle utilizes the technique of parameter shift (Bhuiya & Stanley, 1963) to reduce the value of the minimal function [equation (10)]. The *Shake-and-Bake* method, as implemented in the computer program *SnB* (Weeks & Miller, 1999*a*), has successfully provided *ab initio* solutions for structures containing as many as 1200 independent non-H atoms (Deacon *et al.*, 1998) as well as for large substructures such as the 160-site selenomethionine derivative of ketopantoate hydroxymethyltransferase from *E. coli* (von Delft & Blundell, 2002).

The statistical parameter-shift procedure, a modification of the existing parameter-shift procedure (Chang *et al.*, 1997) implemented in *Shake-and-Bake*, is designed to take advantage of the special properties of the statistical maximal function. The phases are sorted in decreasing order with respect to the values of the associated $|E|$s, and initial values of phases are calculated based on a trial structure having randomly positioned atoms. When considering a given phase ϕ_H, the values of the statistical maximal function [equation (12)] are evaluated respectively with phases $\varphi_H \pm jS$ for $j = 0, 1, \ldots, m$, where S is a predetermined phase shift (*shift size*) and $m = [180/S]$ ($[x]$ is the largest integer whose magnitude does not exceed the magnitude of x). Then the maximum of these values of the statistical maximal function is found, and the phase ϕ_H is updated to reflect that modification. The consideration of ϕ_H is complete, and statistical parameter shift proceeds to the next phase. Refined phase values are used immediately in the subsequent refinement of other phases. The notation STAT-PS(S, k) is used to denote the statistical parameter-shift optimization of the statistical maximal function, using shift size S and k iterations (passes through the phase set) of phase refinement per *Shake-and-Bake* cycle.

Both traditional *Shake-and-Bake* (Weeks & Miller, 1999*a*) and statistical *Shake-and-Bake* were applied to the 12 known centrosymmetric, non-centrosymmetric structures and Se–Met substructures listed in Table 1. For a Se–Met substructure determination, the peak-wavelength anomalous scattering data were used. A sample of 1000 randomly positioned N_μ-atom trial structures (where N_μ is the number of independent atoms in the asymmetric unit) was generated for each data set. For each structure or substructure, the default values of the important, size-dependent, *SnB* parameters are summarized in Table 3. For traditional *Shake-and-Bake*, the default parameter-shift procedure (Weeks & Miller, 1999*b*) was applied and, for statistical *Shake-and-Bake*, the statistical parameter-shift procedure STAT-PS(180°, 1) was used for centrosymmetric structure determination, and STAT-PS(60°, 3) was used for non-centrosymmetric structure and Se–Met substructure determination.

The *success rate* reported in this paper is defined as the percentage of trial structures that go to solution. When performing *post mortem* studies using data for previously known structures, a trial structure subjected to the *Shake-and-Bake* procedure is counted as a solution if there is a close match between the peak positions produced by *Shake-and-Bake* and the true atomic positions for some choice of origin and enantiomorph. Of course, in actual applications to unknown structures, potential solutions are identified on the basis of objective function values. For traditional *Shake-and-Bake*, potential solutions are identified by the values of the cosine minimal function [equation (10)] and, for statistical *Shake-and-Bake*, potential solutions are identified by the values of the statistical maximal function [equation (12)]. Table 4 summarizes success rates obtained from both traditional and statistical *Shake-and-Bake* for the 12 structures listed in Table 1.

4. Conclusions and discussion

The statistical maximal function and its maximal principle have been formulated, and the corresponding statistical *Shake-and-Bake* and its associated statistical parameter-shift procedure have been proposed and tested. For ten out of twelve test structures, the success rate from statistical *Shake-and-Bake* is higher than that of traditional *Shake-and-Bake*. The test results have confirmed that the statistical maximal principle is valid and statistical *Shake-and-Bake* is capable of determining crystal structures including centrosymmetric, non-centrosymmetric and heavy-atom substructures.

The statistical parameter-shift procedure employed in the statistical *Shake-and-Bake* method is just one of many possible reciprocal phase-refinement procedures. Although this procedure has not been optimized, the results based on the success rate show that the statistical *Shake-and-Bake* compares favorably with traditional *Shake-and-Bake*. We believe that the potential of the statistical approach to the phase problem has not yet been fully exploited. It is our intention to optimize statistical *Shake-and-Bake*, particularly the statistical parameter-shift procedure and its parameters, such as shift size and number of iterations, to achieve the maximal success.

The statistical maximal function is less computationally intensive than the cosine minimal function. It does not require the calculation of the trigonometric functions or the Bessel functions. It also does not require the re-calculation of the objective function at several different locations owing to the phase shift during the parameter-shift procedure; once the initial triplet value is calculated, then all triplet locations owing to phase shift are automatically determined, so are their contributions to the maximal function. Despite the reduced computational intensity of the statistical maximal function, the computing time of statistical *SnB* is almost identical to that of traditional *SnB*. Owing to the greater success rate of statistical *Shake-and-Bake*, the average time to solution with statistical *Shake-and-Bake* is less than that of traditional *Shake-and-Bake*.

Finally, the statistical maximal function can be easily converted into a statistical minimal function by means of $1 - S(\phi)$, where $S(\phi)$ is the statistical maximal function defined by equation (12).

This research was supported by NIH grant GM-46733.

References

Alfonso, M. & Stoeckli-Evans, H. (2001). *Acta Cryst.* E**57**, o242–o244.

Arjunan, P., Nemeria, N., Brunskill, A., Chandrasekhar, K., Sax, M., Yan, Y., Jordan, F., Guest, J. R. & Furey, W. (2002). *Biochemistry*, **41**, 5213–5221.

Bhuiya, A. K. & Stanley, E. (1963). *Acta Cryst.* **16**, 981–984.

Camiolo, S., Coles, S. J., Gale, P. A., Hursthouse, M. B. & Paver, M. A. (2001). *Acta Cryst.* E**57**, o258–o260.

Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. A. (1997). *Acta Cryst.* A**53**, 436–444.

Chen, C. C. H., Zhang, H., Kim, A. D., Howard, A., Sheldrick, G. M., Mariano-Dunnaway, D. & Herzberg, O. (2002). *Biochemistry*, **41**, 13162–13169.

Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.

Deacon, A. M., Weeks, C. M., Miller, R. & Ealick S. E. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.

Delft, F. von & Blundell, T. L. (2002). *Acta Cryst.* A**58** (Supplement), C239.

DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.

Hauptman, H. A. & Karle, J. (1953). *Am. Monograph* No. 3. *Solution of the Phase Problem. I. The Centrosymmetric Crystal.* Michigan: American Crystallographic Association.

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.

Karle, J. & Hauptman, H. A. (1956). *Acta Cryst.* **9**, 635–651.

McCarthy, A. A., Baker, H. M., Shewry, S. C., Patchett, M. L. & Baker, E. N. (2001). *Structure*, **9**, 637–646.

Marshall, G. R., Hodgkin, E. E., Langs, D. A., Smith, G. D., Zabrocki, J. & Leplawy, M. T. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 487–491.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.

Ohba, S., Hiratsuka, T. & Tanaka, K. (2002). *Acta Cryst.* E**58**, o1013–o1015.

Pletnev, V. Z., Galitskii, N. M., Smith, G. D., Weeks, C. M. & Duax, W. L. (1980). *Biopolymers*, **19**, 1517–1534.

Pletnev, V. Z., Ivanov, V. T., Langs, D. A., Strong, P. & Duax, W. L. (1992). *Biopolymers*, **32**, 819–827.

Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.

Turner, M. A., Yuan, C. S., Borchardt, R. T., Hershfield, M. S., Smith, G. D. & Howell, P. L. (1998). *Nature Struct. Biol.* **5**, 369–376.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.

Weeks, C. M. & Miller, R. (1999*a*). *J. Appl. Cryst.* **32**, 120–124.

Weeks, C. M. & Miller, R. (1999*b*). *Acta Cryst.* D**55**, 492–500.

Zhuang, J. P., Zheng, Y. & Zhang, W. Q. (2002). *Acta Cryst.* E**58**, o720–o722.