Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

σ_R^2, a reciprocal-space measure of the quality of macromolecular electron-density maps

Thomas C. Terwilliger

Structural Biology Group, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Correspondence e-mail: terwilliger@lanl.gov

Received 9 November 1998
Accepted 2 March 1999

It has previously been shown that the presence of distinct regions of solvent and protein in macromolecular crystals leads to a high value of the standard deviation of local r.m.s. electron density and that this can in turn be used as a reliable measure of the quality of macromolecular electron-density maps [Terwilliger & Berendzen (1999*a*). *Acta Cryst.* D**55**, 501–505]. Here, it is demonstrated that a similar measure, σ_R^2, the variance of the local roughness of the electron density, can be calculated in reciprocal space. The formulation is suitable for rapid evaluation of macromolecular crystallographic phases, for phase improvement and for *ab initio* phasing procedures.

1. Introduction

A key step in the determination of macromolecular crystal structures, either by direct methods or by more traditional MAD or MIR approaches, is the evaluation of the quality of an electron-density map. In applying direct methods to macromolecular crystal structure determination, statistical relationships derived from characteristics of small-molecule structures (*e.g.* Sheldrick, 1990; Weeks *et al.*, 1995; Hauptman, 1997) are typically used to discriminate between possible phase sets. In the MAD or MIR approaches, the crystallographer typically manually examines an electron-density map and equates its interpretability with its quality. There would be considerable utility in having objective measures of the quality of electron-density maps which include as many features of macromolecular crystals as possible. Such measures could be used to choose between possible phase sets in *ab initio* methods and between possible heavy-atom solutions in the MIR and MAD methods. Additionally, if the measure of quality could be expressed in a simple reciprocal-space formulation, the measure could be used to improve phase quality or even to determine phases *ab initio*.

One measure of the quality of macromolecular electron-density maps which has been proposed is an automated analysis of the connectivity of electron-density maps (Baker *et al.*, 1993). This approach works well for evaluation of a map, but unfortunately it has proven difficult to use in phase improvement. We have recently demonstrated that an evaluation of the distinction between solvent and protein regions can be a very powerful criterion for scoring electron-density maps (Terwilliger & Berendzen, 1999*a*,*b*). Our approach is based on the well known observation that macromolecular crystals typically contain distinct regions of protein (where the local variation of electron density from point to point is very high) and solvent (where the electron density is essentially constant). This observation has been the basis of widely used solvent-flattening procedures (Wang,

1985; Xiang *et al.*, 1993; Podjarny *et al.*, 1987; Abrahams *et al.*, 1994; Zhang & Main, 1990).

We have used the difference between protein and solvent regions to generate an objective measure of the quality of a macromolecular electron-density map. Firstly, we calculated the local r.m.s. electron density near each grid point in the asymmetric unit, omitting the F_{000} term in the Fourier synthesis. In this way, the local r.m.s. density is very small in the solvent region but large in the protein region. We then determined the standard deviation of this local r.m.s. density over the entire asymmetric unit and use it as a figure of merit of the phasing. Maps which have a uniform distribution of local r.m.s. density have low values of the standard deviation; those with distinct protein and solvent regions have higher values. We have found this measure very useful in differentiating between heavy-atom solutions in the MIR and MAD approaches, as well as in identification of the hand of heavy-atom solutions when anomalous differences have been measured (Terwilliger & Berendzen, 1999*a*).

Although it is difficult to express the standard deviation of local r.m.s. electron density in a reciprocal-space formulation, a very closely related characteristic, the variance of the local roughness, can be calculated readily. Here, we define this variance of the local roughness as the overall variance of the local variance of electron density in a map, and show how it can be calculated in reciprocal space. The expression we derive is suitable as a figure of merit for phase-quality evaluation, for phase improvement and for *ab initio* phasing methods.

2. Theory

In our previous work, we calculated the standard deviation of local r.m.s. electron density in a map. It was calculated using a grid with spacing approximately one-third of the resolution of the map in boxes five grid units on an edge, and the standard deviation of the local r.m.s. density was obtained from overlapping boxes throughout the asymmetric unit of the crystal (Terwilliger & Berendzen, 1999*a*). We found that the precise size and overlap of the boxes had only small effects on the calculation. Here, we use a closely related but more generalizable approach, in which the overall variance of the local roughness of electron density is calculated. Instead of using overlapping boxes to determine the variation of local mean-square density from point to point in the cell, we use a windowing function to define the region over which the local variance (roughness) of electron density is calculated. Any windowing function could be used for this purpose, but a particularly convenient one is a Gaussian function.

The local roughness in a map $[r(\mathbf{x})]$ can be represented by the weighted variance of electron density in a region defined by a windowing function centered at $\mathbf{x}$:

$$r(\mathbf{x}) = \int_{R^3} [\rho(\mathbf{x}') - \overline{\rho}(\mathbf{x})]^2 g(\mathbf{x} - \mathbf{x}') \mathrm{d}^3\mathbf{x}', \quad (1)$$

or equivalently

$$r(\mathbf{x}) = \int_{R^3} \rho^2(\mathbf{x}') g(\mathbf{x} - \mathbf{x}') \mathrm{d}^3\mathbf{x}' - \overline{\rho}(\mathbf{x})^2, \quad (2)$$

where $\overline{\rho}(\mathbf{x})$ is the mean local electron density, given by

$$\overline{\rho}(\mathbf{x}) = \int_{R^3} \rho(\mathbf{x}') g(\mathbf{x} - \mathbf{x}') \mathrm{d}^3\mathbf{x}', \quad (3)$$

and $g(\mathbf{x})$ is an arbitrary windowing function. If the windowing function is a three-dimensional Gaussian function with unit volume and a variance (for each of the components x, y, z) of σ^2 then it can be expressed as

$$g(\mathbf{x}) = (1/2\pi)^{1/2} (1/\sigma^3) \exp[-0.5(||\mathbf{x}||^2/\sigma^2)]. \quad (4)$$

The variance (σ_R^2) of this local roughness of electron density over the entire unit cell is then given by

$$\sigma_R^2 = (1/V) \int_V r^2(\mathbf{x}) \mathrm{d}^3\mathbf{x} - \bar{r}^2, \quad (5)$$

where $\bar{r} = (1/V) \int r(\mathbf{x})$ and V is the volume of the unit cell.

To calculate the variance of local roughness of the electron density, σ_R^2, in reciprocal space, we use the facts that the first term on the right-hand side of (2) represents the convolution of $\rho^2(\mathbf{x})$ and $g(\mathbf{x})$, and that $\overline{\rho}(\mathbf{x})$ in (2) is in turn the convolution of $\rho(\mathbf{x})$ and $g(\mathbf{x})$. The electron density $\rho(\mathbf{x})$, assumed to be a real function, and the squared electron density $\rho^2(\mathbf{x})$ can be expressed as (*cf.* Bracewell, 1986)

$$\rho(\mathbf{x}) = \sum_{\mathbf{h}} \mathbf{F}_{\mathbf{h}} \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}), \quad (6)$$

and

$$\rho^2(\mathbf{x}) = \sum_{h} \mathbf{B}_{\mathbf{h}} \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}), \quad (7)$$

respectively, where $\mathbf{h} \equiv (h\mathbf{a}^*, k\mathbf{b}^*, l\mathbf{c}^*)$ and the reciprocal lattice vectors are $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$. The coefficients $\mathbf{B}_{\mathbf{h}}$ can be calculated from the structure factors $\mathbf{F}_{\mathbf{h}}$ using the relation

$$\mathbf{B}_{\mathbf{h}} = \sum_{\mathbf{k}} \mathbf{F}_{\mathbf{k}} \mathbf{F}_{\mathbf{h}-\mathbf{k}}, \quad (8)$$

summing over all values of $\mathbf{k}$. The Gaussian function $g(\mathbf{x})$ can be readily expressed in Fourier space; it appears as the temperature factor in the Fourier transform of a Gaussian distribution of electron density about an atom, for example. An expression for a Gaussian centered at the origin with unit volume and a variance of ρ^2 is

$$g(\mathbf{x}) = \sum_{\mathbf{h}} \mathbf{G}_{\mathbf{h}} \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}), \quad (9)$$

where

$$\mathbf{G}_{\mathbf{h}} = \exp(-2\sigma^2 \pi^2 S_{\mathbf{h}}^2) \quad (10)$$

and $S_{\mathbf{h}}$ is the magnitude of the scattering vector $||\mathbf{h}|| = 2 \sin\theta/\lambda$.

As $\overline{\rho}(\mathbf{x})$ (3) is the convolution of $\rho(\mathbf{x})$ and $g(\mathbf{x})$, we can write

$$\overline{\rho}(\mathbf{x}) = \sum_{\mathbf{h}} \mathbf{Q}_{\mathbf{h}} \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}), \quad (11)$$

where the coefficients $\mathbf{Q}_{\mathbf{h}}$ are simply the original structure factors $\mathbf{F}_{\mathbf{h}}$ damped by the exponential factors $\mathbf{G}_{\mathbf{h}}$,

$$\mathbf{Q}_{\mathbf{h}} = \mathbf{F}_{\mathbf{h}} \mathbf{G}_{\mathbf{h}}. \quad (12)$$

The second term on the right-hand side of (2) can now be expressed using (7) and (8) as

$$\overline{\rho}(\mathbf{x})^2 = \sum_{\mathbf{h}} \mathbf{B}_{\mathbf{h}}^{\mathrm{AVG}} \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}), \tag{13}$$

where the coefficients $\mathbf{B}_{\mathbf{h}}^{\mathrm{AVG}}$ are based on the dampened structure factors $\mathbf{Q}_{\mathbf{k}}$ in (12),

$$\mathbf{B}_{\mathbf{h}}^{\mathrm{AVG}} = \sum_{\mathbf{k}} \mathbf{Q}_{\mathbf{k}}\mathbf{Q}_{\mathbf{h}-\mathbf{k}}. \tag{14}$$

Next, as the first term on the right-hand side of (2) is the convolution of $\rho^2(\mathbf{x})$ and $g(\mathbf{x})$, we can write

$$\int_{R^3} \rho^2(\mathbf{x}')g(\mathbf{x}-\mathbf{x}')\mathrm{d}^3\mathbf{x}' = \sum_{\mathbf{h}} \mathbf{T}_{\mathbf{h}} \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}), \tag{15}$$

where the coefficients $\mathbf{T}_{\mathbf{h}}$ are given by

$$\mathbf{T}_{\mathbf{h}} = \mathbf{B}_{\mathbf{h}}\mathbf{G}_{\mathbf{h}}. \tag{16}$$

We can now express the local roughness of a map (1) in the form

$$r(\mathbf{x}) = \sum_{\mathbf{h}} \mathbf{R}_{\mathbf{h}} \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}), \tag{17}$$

where the coefficients $\mathbf{R}_{\mathbf{h}}$ are given by

$$\mathbf{R}_{\mathbf{h}} = \mathbf{B}_{\mathbf{h}}\mathbf{G}_{\mathbf{h}} - \mathbf{B}_{\mathbf{h}}^{\mathrm{AVG}}. \tag{18}$$

The desired variance σ_R^2 in (5) is composed of two parts, the mean value of $r^2(\mathbf{x})$ and the square of the mean value of $r(\mathbf{x})$ over the unit cell. The mean value of $r(\mathbf{x})$ over the unit cell is simply the $\mathbf{h}$ = (0, 0, 0) term of its corresponding transform, $\mathbf{R}_{000}$. Similarly, the mean value of $r^2(\mathbf{x})$ is given by the $\mathbf{h}$ = (0, 0, 0) term of its transform. Using Parseval's theorem (*cf.* Bracewell, 1986), the mean value of $r^2(\mathbf{x})$ can be expressed in the form

$$(1/V)\int_V r^2(\mathbf{x}) = \sum_{\mathbf{h}} ||\mathbf{R}_{\mathbf{h}}||^2, \tag{19}$$

where the integral is taken over the unit-cell volume.

Finally, the variance of local roughness (σ_R^2) in (5) can be written as

$$\sigma_R^2 = \sum_{\mathbf{h}} ||\mathbf{R}_{\mathbf{h}}||^2 - \mathbf{R}_{000}^2 \tag{20}$$

or more simply as

$$\sum_{\mathbf{h}\neq(000)} ||\mathbf{R}_{\mathbf{h}}||^2. \tag{21}$$

3. Discussion

(21) is a representation in reciprocal space of σ_R^2, the variance of the local roughness of electron density in a Fourier synthesis. In the case of macromolecular crystals containing well defined regions of protein and solvent, this variance tends to be very high, as protein-containing areas of the unit cell are very rough and solvent-containing areas are very smooth (Terwilliger & Berendzen, 1999*a*). Consequently, the value of this variance can be used as a measure of the relative qualities of various possible phase sets for a macromolecular structure.

The variance of local roughness, σ_R^2, in (21) is given by the sum of squares of the coefficients $\mathbf{R}_{\mathbf{h}}$, other than $\mathbf{R}_{000}$, in the Fourier synthesis for the local roughness, $r(\mathbf{x})$. This is equivalent to noting that σ_R^2 is simply the overall mean square value of the local roughness, after subtracting the overall average value of $\mathbf{R}_{000}$. The coefficients $\mathbf{R}_{\mathbf{h}}$ for the local roughness, given in (18), each contain two terms, $\mathbf{B}_{\mathbf{h}}\mathbf{G}_{\mathbf{h}}$ and $\mathbf{B}_{\mathbf{h}}^{\mathrm{AVG}}$. The first term, $\mathbf{B}_{\mathbf{h}}\mathbf{G}_{\mathbf{h}}$, consists of coefficients in the Fourier series expression (15) for the local mean-square electron density. The second term, $\mathbf{B}_{\mathbf{h}}^{\mathrm{AVG}}$, are coefficients in the Fourier series expression for the local mean electron density, squared. The difference corresponds to the local variance of the electron density, which we describe as local roughness.

An important feature of (21) is that only the low-order terms are large. This is a consequence of the presence of the exponential terms $\mathbf{G}_{\mathbf{h}}$ multiplying the $\mathbf{B}_{\mathbf{h}}$ terms in (18) and multiplying the $\mathbf{F}_{\mathbf{h}}$ terms in (12). Because of this, σ_R^2 in (21) is, to a first approximation, the sum of the squares of the lowest-order terms in the Fourier series (7) describing $\rho^2(\mathbf{x})$. The magnitudes of these low-order terms describe how well defined the regions of the unit cell are which contain low and high values of $\rho^2(\mathbf{x})$. If the distribution of $\rho^2(\mathbf{x})$ is relatively uniform in the unit cell, then the low-order terms in this Fourier series will be small. If the distribution is highly non-uniform then the low-order terms, and hence σ_R^2, will be large.

(21) has several important properties which should be emphasized. The most significant is that the exponential term limits the range of $\mathbf{h}$ over which the terms in the summation are large to those with small $||\mathbf{h}||$. This means that evaluating σ_R^2 can be rapid. The calculation of each $\mathbf{B}_{\mathbf{h}}$ in (8) or $\mathbf{B}_{\mathbf{h}}^{\mathrm{AVG}}$ in (14) requires just one pass through all reflections. As only small values of $\mathbf{h}$ make a large contribution to σ_R^2, a relatively small number of passes through the reflections are necessary to calculate σ_R^2. The potential rapidity of calculation of σ_R^2 means that Monte Carlo methods or methods based on the genetic algorithm could potentially be used to optimize σ_R^2 even in cases with large numbers of reflections. If a windowing function other than a Gaussian is used, or if the Gaussian function has a very narrow width, however, the number of terms needed to accurately evaluate σ_R^2 would not necessarily be small. In general, the calculation of σ_R^2 using the low-order terms in (21) corresponds to truncation of the spectrum of the windowing function at some resolution.

The second significant aspect of (21) is that the value of σ_R^2 depends on the crystallographic phases in an easily calculable way. It is straightforward to differentiate (21) with respect to individual phases. This means that matrix methods can be used to adjust the phases to maximize σ_R^2. As reflections only interact significantly in (8) with other reflections which differ in $\mathbf{k}$ by a small number, such matrix methods would have to involve at most only a fraction of the elements in the matrix and possibly just diagonal elements. This kind of approach could be used to combine the maximization of σ_R^2 with that of other direct-methods figures of merit to improve the ability of direct-methods to solve macromolecular structures.

As (21) is essentially a reciprocal-space formulation of the real-space measure of map quality which we have already examined in detail (Terwilliger & Berendzen, 1999*a*), most of the properties of the two formulations will be very similar. In Fig. 1, we present a set of model calculations using (21) to evaluate electron-density maps in reciprocal space. 6200 model data from 20 to 3.0 Å were generated based on coordinates from a dehalogenase enzyme from *Rhodococcus* species ATCC 55388 (American Type Culture Collection, 1992) determined recently in our laboratory. The protein contains 316 amino-acid residues and crystallizes in space group $P2_12_12$ with unit-cell dimensions $a = 94$, $b = 80$, $c = 43$ Å and one molecule in the asymmetric unit (J. Newman, personal communication). Fig. 1(*a*) shows results for a total of 2000 phase sets generated from the model data, with phase errors ranging from 0–150°. These model data sets were analyzed using (21) with a value of $\sigma = 6$ Å and including all 364 terms for which the exponential term G(**h**) in (10) has a value of 0.0001 or larger. The logarithm of the variance in local roughness, $\log(\sigma_R^2)$, is plotted in Fig. 1(*a*) as a function of the cosine of the mean phase error in the data set. For phase sets with $\langle\cos(\Delta\theta)\rangle$ of $\sim$0.3 or greater, the logarithm of the variance in local roughness is quite closely related to the phase accuracy. For phase sets with lower $\langle\cos(\Delta\theta)\rangle$, there is only a small correlation.

Fig. 1(*b*) shows the practical implications of the data in Fig. 1(*a*) and also illustrates that only low-order terms in (21) are necessary for calculating σ_R^2. In Fig. 1(*b*), the data in Fig. 1(*a*) are analyzed to estimate the probability that a correct choice of the better of two phase sets can be determined from the logarithm of the variance of local roughness. Fig. 1(*b*) shows analyses of four groups of 2000 phase sets each. In each of the four analyses, a different minimum value of the exponential term G(**h**) was used, ranging from 0.0001 to 0.1. To obtain Fig. 1(*b*), the data in Fig. 1(*a*) were grouped into pairs of sets differing by 0.1 ± 0.05 units in $\langle\cos(\Delta\theta)\rangle$. Each member of each set was compared with each member of the paired set, and the fraction of times that the member with the higher value of $\log(\sigma_R^2)$ also had the higher value of $\langle\cos(\Delta\theta)\rangle$ was plotted.

Fig. 1(*b*) shows that, as expected in phase sets with very low phase accuracy ($\langle\cos(\Delta\theta)\rangle < 0.25$), the value of $\log(\sigma_R^2)$ leads to only a 50% chance of choosing the better of two phase sets which differ in accuracy. For phase sets with values of $\langle\cos(\Delta\theta)\rangle$ from 0.25 to 0.4, however, the probability of choosing the better of two phase sets differing by this amount increases from 0.6 to 0.9. The 58 lowest order terms in the series in (21) give almost the same likelihood of making a correct choice as the 364 lowest order terms. This means that high-order terms can be ignored without a substantial effect.

4. Conclusions

The reciprocal-space formulation presented here has major advantages compared with the real-space calculations carried out previously (Terwilliger & Berendzen, 1999*a*). These are that the variance σ_R^2 can be calculated without a Fourier transform and that potentially phases can be adjusted to maximize the variance. The rapid calculation of variance means that it can be used as a measure of the quality of phases in many different trials, and the potential for maximization of the variance means that it can be used in phase improvement and possibly even *ab initio* phasing algorithms. The most powerful means for phase improvement for macromolecules without non-crystallographic symmetry is at present solvent flattening (Wang, 1985; Xiang *et al.*, 1993; Podjarny *et al.*, 1987; Abrahams *et al.*, 1994; Zhang & Main, 1990). Carrying out this type of procedure requires that the electron-density map be of sufficiently high quality that an envelope defining the boundary between protein and solvent can be reliably calculated. (21) provides a means for improving phases even before the boundary is clearly defined. Maximizing σ_R^2 will maximize the distinction between protein and solvent regions without requiring a knowledge of where each are located. Consequently, (21) may be useful in cases where solvent flattening is not effective, as well as providing a complementary approach in cases where phases are good to begin with.

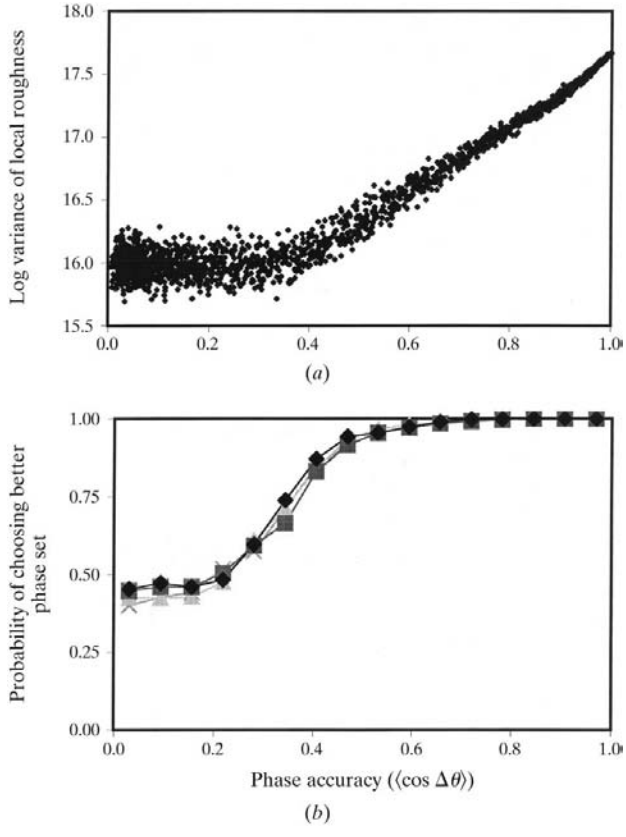

Figure 1
Calculation of variance of local roughness using (21). (*a*) The logarithm of σ_R^2 is plotted for 2000 model phase sets, as described in the text. The abscissa is $\langle\cos(\Delta\theta)\rangle$, the mean value of the effective figure of merit of the phase set. (*b*) The probability of choosing the better of two phase sets which differ in quality by 0.1 units of $\langle\cos(\Delta\theta)\rangle$ is plotted for model data obtained as in (*a*), using the 364 lowest order terms (diamonds), 249 lowest order terms (triangles), 145 lowest order terms (squares) or 58 lowest order terms (crosses), as described in the text.

There are several aspects of the reciprocal-space formulation which remain to be optimized. One is the choice of the windowing function. We have chosen a Gaussian function, but the derivation we carried out is independent of the windowing function and any function could have been used. A Gaussian is particularly convenient because it results in strongly damped coefficients that become very small for all but small values of $|\mathbf{h}|$. Other windowing functions, however, might yield better measures of the quality of the electron-density map, and a survey of other functions might improve the algorithm. Another possibility might be to construct a histogram of values of σ_R^2 from many solved protein structures which could in turn be used to construct a data-likelihood model for estimation of phase errors. Such an approach could be considerably more powerful than the one described here because it would give probability information which could be combined in a Bayesian approach with other sources of phase information.

The author is grateful for discussion with Randy Read, Joel Berendzen and Janet Newman, for exceptionally helpful comments from anonymous reviewers and for support from the National Institutes of Health and the Department of Energy.

References

Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. (1994). *Nature (London)*, **370**, 621–628.
American Type Culture Collection (1992). *Catalogue of Bacteria and Bacteriophages*, 18th ed., pp. 271–272.
Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 186–192.
Bracewell, R. N. (1986). *The Fourier Transform and Its Applications.* New York: McGraw-Hill.
Hauptman, H. (1997). *Curr. Opin. Struct. Biol.* **7**, 672–680.
Podjarny, A. D., Bhat, T. N. & Zwick, M. (1987). *Annu. Rev. Biophys. Biophys. Chem.* **16**, 351–373.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Terwilliger, T. C. & Berendzen, J. (1999*a*). *Acta Cryst.* D**55**, 501–505.
Terwilliger, T. C. & Berendzen, J. (1999*b*). *Acta Cryst.* D**55**, 849–861.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Weeks, C. M., Hauptman, H. A., Smith, G. D. & Blessing, R. H. (1995). *Acta Cryst.* D**51**, 33–38.
Xiang, S., Carter, C. W. Jr, Bricogne, G. & Gilmore, C. J. (1993). *Acta Cryst.* D**49**, 193–212.
Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst.* A**46**, 41–46.