Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Evaluation of phase accuracy *via* topological and geometrical analysis of electron-density maps

Christos Colovos,[a] Eric A. Toth[a,b] and Todd O. Yeates[a]*

[a]Department of Chemistry and Biochemistry, DOE-MBI Laboratory of Structural Biology and Molecular Medicine, Box 951570, University of California, Los Angeles, CA 90095-1570, USA, and [b]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, USA

Correspondence e-mail: yeates@mbi.ucla.edu

Received 3 May 2000
Accepted 9 August 2000

An empirical function is developed to measure the protein-like character of electron-density maps. The function is based upon a systematic analysis of numerous local and global map properties or descriptors. Local descriptors measure the occurrence throughout the unit cell of unique patterns on various defined templates, while global descriptors enumerate topological characteristics that define the connectivity and complexity of electron-density isosurfaces. We examine how these quantitative descriptors vary as error is introduced into the phase sets used to generate maps. Informative descriptors are combined in an optimal fashion to arrive at a predictive function. When the topological and geometrical analysis is applied to protein maps generated from phase sets with varying amounts of error, the function is able to estimate changes in average phase error with an accuracy of better than 10°. Additionally, when used to monitor maps generated with experimental phases from different heavy-atom models, the analysis clearly distinguishes between the correct heavy-atom substructure solution and incorrect heavy-atom solutions. The function is also evaluated as a tool to monitor changes in map quality and phase error before and after density-modification procedures.

1. Introduction

Initial phase information from a crystallographic experiment often results in an electron-density map of insufficient quality to determine a protein structure. By adjusting the phases with additional experiments or with computational methods, the quality of the resultant map can be enhanced. Several computational methods exploit characteristic properties of accurate electron-density maps to improve the phases on which the maps rely. Popular methods include solvent flattening (Wang, 1985), skeletonization (Baker *et al.*, 1993), histogram matching (Zhang & Main, 1990) and non-crystallographic symmetry averaging (Bricogne, 1974). The constraints employed by these methods are useful in improving phases, but they do not fully utilize the unique properties of well resolved protein electron-density maps.

Previous work in the area of density modification has focused mainly on restraining electron-density maps to obey relatively simple properties that can be anticipated easily. For example, solvent flattening takes advantage of the expectation that the solvent regions will have uniform electron density. Likewise, direct methods for phasing are based on relatively simple constraints of positivity and atomicity of electron density (Hauptman, 1986; Karle, 1986). However, because proteins are very complex molecules, one might expect their electron-density maps to obey many complex properties that may not be anticipated without a systematic study. By visual

inspection, a skilled observer can identify global and local features indicative of a protein structure, but to date few automated computational approaches have been developed to analyze complex map properties. Two-dimensional histograms of electron density have been used to evaluate protein map quality and phase error (Goldstein & Zhang, 1998). Ioerger *et al.* (1999) recently described a pattern-recognition algorithm to interpret electron-density maps in an effort to automatically build a protein model. Studies by Terwilliger & Berendzen (1999) have shown that local deviations of electron density in a map give a good indication of map quality and have suggested potential applications in evaluating the quality of phase sets. A more complete and systematic enumeration of complex properties obeyed by accurate protein electron-density maps could bring new forces to bear on the problems of estimating and improving diffraction phases.

The method presented here is based on the premise that better phases produce electron-density maps with more protein-like features. Under this assumption, a method that quantifies the appropriate properties of an electron-density map should be able to give an estimate of the accuracy of the corresponding phase set. By comparing maps, a function based on this type of analysis could discriminate between manipulations that improve or degrade phase accuracy. This capability could be used to evaluate candidate heavy-atom models, the results of density-modification procedures and *ab initio* phase calculations.

In this paper, we present a new method to evaluate the protein-like appearance of electron-density maps based on a topological and geometrical analysis (TGA). The method is based on a systematic search for calculable properties of electron-density maps that correlate with phase error. Many of the map properties or descriptors examined are informative and combine to give an empirical function that quantifies the overall accuracy of diffraction phase sets. We show how the function can be used to calculate the relative error in various phase sets, to identify heavy-atom substructure solutions and to evaluate the progress of phase refinement by density modification.

2. Methods

2.1. Contents of the database and generation of voxel maps

A small database of representative protein structures was created from the Protein Data Bank (Sussman *et al.*, 1998). Attention was restricted to structures with a resolution of 2.0 Å or better and an R factor less than 20%. From these, we selected nine protein structures without extraordinarily large unit-cell dimensions spanning several different crystal systems (Table 1). From the coordinates, we calculated structure factors and phases to 3.0 Å resolution using the program *SFALL* (Collaborative Computational Project, Number 4, 1994).

In order to generate a multitude of phase sets with known amounts of phase error, random errors were added to the 'correct' (model) phase sets. Phase sets were generated with average errors ranging from 0 to 90° in 10° increments. In each phase set, error was introduced as follows. Within each of five resolution shells, we added a random (uniformly distributed) phase error to each reflection. In order to roughly simulate the dependence of phase error on resolution, the average overall phase error was multiplied in each shell by a factor ranging from 1.4 at the highest resolution to 0.6 at the lowest resolution. At every level of overall phase error, 25 randomizations were performed to give 25 different phase sets with similar overall error. The collection of phase sets was used to calculate a large family of electron-density maps with known phase errors; 450 for each of the nine proteins.

Table 1
Database structures.

PDB code	Space group	Unit-cell parameters (Å,°)
1mbw	$P6$	$a = b = 91.2$, $c = 45.8$, $\alpha = \beta = 90.0$, $\gamma = 120.0$
1ova	$P1$	$a = 62.9$, $b = 84.7$, $c = 71.5$, $\alpha = 87.5$, $\beta = 104.0$, $\gamma = 108.5$
2fcr	$P2_12_12_1$	$a = 63.6$, $b = 48.8$, $c = 56.8$, $\alpha = \beta = \gamma = 90.0$
2ltn	$P2_12_12_1$	$a = 50.7$, $b = 61.2$, $c = 136.6$, $\alpha = \beta = \gamma = 90.0$
2lym	$P4_32_12$	$a = b = 79.2$, $c = 38.0$, $\alpha = \beta = \gamma = 90.0$
2lzt	$P1$	$a = 28.3$, $b = 32.0$, $c = 34.3$, $\alpha = 88.5$, $\beta = 108.6$, $\gamma = 111.9$
2plt	$P3_2$	$a = b = 61.8$, $c = 25.2$, $\alpha = \beta = 90.0$, $\gamma = 120.0$
3cla	$R32$	$a = b = 107.6$, $c = 123.6$, $\alpha = \beta = 90.0$, $\gamma = 120.0$
5cpa	$P2_1$	$a = 51.6$, $b = 60.3$, $c = 47.3$, $\alpha = \gamma = 90.0$, $\beta = 97.3$

Maps were calculated using *FFT* from the *CCP*4 package (Collaborative Computational Project, Number 4, 1994), normalized using the program *AL-MAPMAN* (Kleywegt & Jones, 1996) and represented as voxel maps, as described

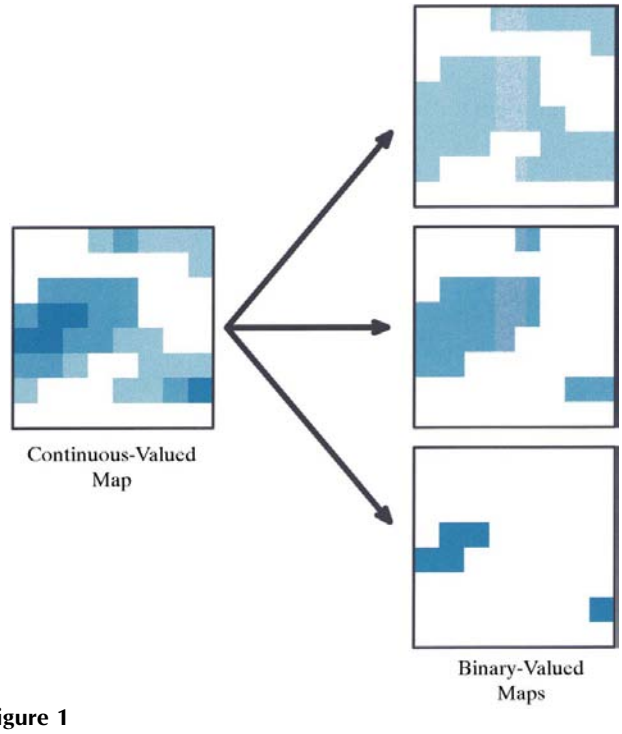

Figure 1
Decomposition of a continuous-valued electron-density map into a family of binary maps. Strong, medium and weak levels of electron density are shown in shades of blue. In a binary map, each grid point is either above or below the chosen electron-density contour level.

below. In all cases, diffraction data were included to 3 Å resolution and electron-density maps were calculated at a spacing of 1 Å, as nearly as possible. Thus, each volume element (voxel) is of the order of 1 Å on an edge.

Our subsequent analyses are based on binary-valued electron-density maps. At a given electron-density contour level (*e.g.* 0.4σ), grid points that are above the cutoff are given a value of 1, while those below it are given a value of 0. A binary assignment of density does not completely reflect the diverse and complex features of electron-density maps. In order to model the information contained in continuous-valued maps, we generate a series of binary maps from each continuous-valued map by choosing different contour levels (Fig. 1). In the calculation presented here, eight different cutoff values were used to generate eight binary maps from each continuous-valued map. The eight contour levels were chosen to be 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6 and 1.8 standard deviations above the mean. Some of these maps represent volumes smaller than the true protein, while others represent volumes which are larger. Since the maps derive from inaccurate phase sets, a wide range of contour levels could be informative. The present scheme for choosing contour levels is only one of many that can be employed and is not necessarily the optimal choice.

2.2. Topological map descriptors

Some of the characteristic properties of protein-like electron-density maps can be quantified by analyzing contour surfaces. Using the reduced representation described in the previous section, we can consider an electron-density map as a collection of volume elements, or voxels, with each voxel encompassing one grid point (~1 Å on an edge). A particular contour level defines a collection of polyhedral surfaces that separates those voxels above the contour level from those below. The topological properties of these polyhedral surfaces can be evaluated in a straightforward fashion.

Topology provides for a description of the complexity of a surface. The number of holes or 'handles' formed by a surface is a measure of its complexity. A surface with no handles (*e.g.* a sphere) is said to have genus 0. A surface with one handle (a torus) is said to have genus 1 and so on (Figs. 2*a* and 2*b*).

Euler's formula describes the topological complexity of a polyhedral surface in terms of its vertices, faces and edges. For a simple three-dimensional object (genus = 0), Euler's formula is

$$V + F - E = 2, \tag{1}$$

where V, F and E represent the number of vertices, faces and edges of the surface. For complex surfaces,

$$V + F - E = \chi, \tag{2}$$

where χ is a topological invariant known as the Euler characteristic. All surfaces with a particular Euler characteristic have the same genus and are topologically equivalent (*i.e.* a cube is topologically equivalent to sphere by a process of continuous deformation).

The genus is related to the Euler characteristic by

$$g = 1 - \chi/2. \tag{3}$$

Therefore, Euler's formula for the genus of a polyhedral surface is

$$V + F - E = 2(1 - g). \tag{4}$$

When examining electron-density maps, we often deal with collections of disjoint surfaces, each of which has an associated genus. To calculate the total topological complexity (denoted here as G_T) for a collection of disjoint surfaces, we note that the genus of each surface, g_i, is given by

$$g_i = -\left[\frac{(V_i + F_i - E_i)}{2}\right] + 1. \tag{5}$$

This leads to an equation for the total number of handles (G_T) in a set of disjoint polyhedral surfaces

$$G_T = \sum_i^n g_i = -\left[\frac{(V + F - E)}{2}\right] + n, \tag{6}$$

where n is the number of surfaces and V, F and E are the total numbers of vertices, faces and edges in the contoured binary map.

When dealing with complex surfaces on a grid, one might employ a variety of schemes to define the connectivity between polyhedral volume elements. We employed two schemes. The first requires two polyhedral surfaces to meet face-to-face in order to be connected. The second allows volume elements to meet face-to-face, edge-to-edge or vertex-to-vertex and still be considered connected (Fig. 3). The two scenarios lead to slightly different numerical results. For each

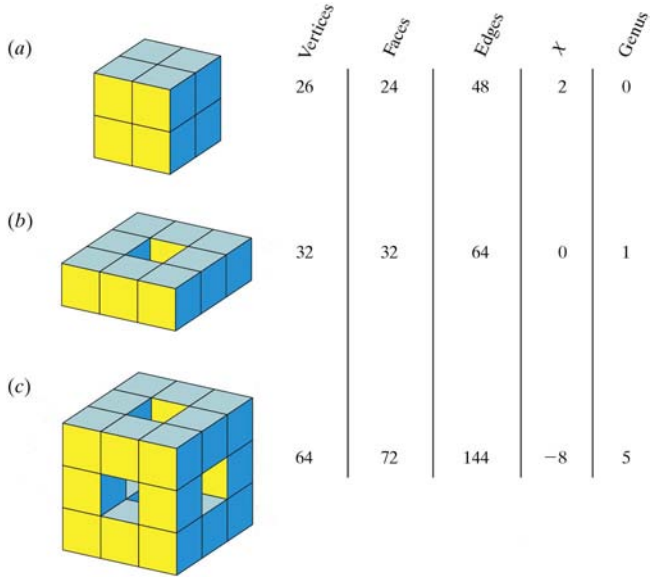

Figure 2
The Euler characteristic as a measure of topological complexity. Three different objects are described: (*a*) a cube, (*b*) a torus and (*c*) a more complex object. The Euler characteristic (χ) is calculated by the formula $\chi = V + F - E$, where V is the number of vertices, F is the number of faces and E is the number of edges. The genus of an object is given by the formula $g = 1 - \chi/2$ and is directly related to the number of handles.

of the two definitions of connectivity, we calculate seven different topological descriptors and the surface area to volume ratio. These descriptors are as follows.

(i) Total number of surfaces enclosing regions of density above the cutoff level (npos).

(ii) Total number of surfaces enclosing regions of density below the cutoff level (nneg).

(iii) Number of disjoint surfaces in the electron-density map (nsurf).

(iv) Euler characteristic (χ) of the electron-density surfaces (2).

(v) Total complexity (G_T) of the electron-density surfaces (6).

(vi) The largest continuous volume of density above the cutoff level (surfmax).

(vii) The largest continuous volume of density below the cutoff level (solvmax).

(viii) Surface area to volume ratio (SA/V).

2.3. Geometric map descriptors

Topology describes global properties. In contrast, if we focus on local regions, the evaluation of binary-valued electron-density maps becomes an exercise in geometric pattern recognition. We expect a map derived from a high-quality phase set to have a noticeably different distribution of patterns than any map derived from a poor phase set. If a statistically significant difference in the distribution of patterns exists between maps generated from accurate and inaccurate phase sets, then an algorithm that measures these differences could quantify the accuracies of various phase sets.

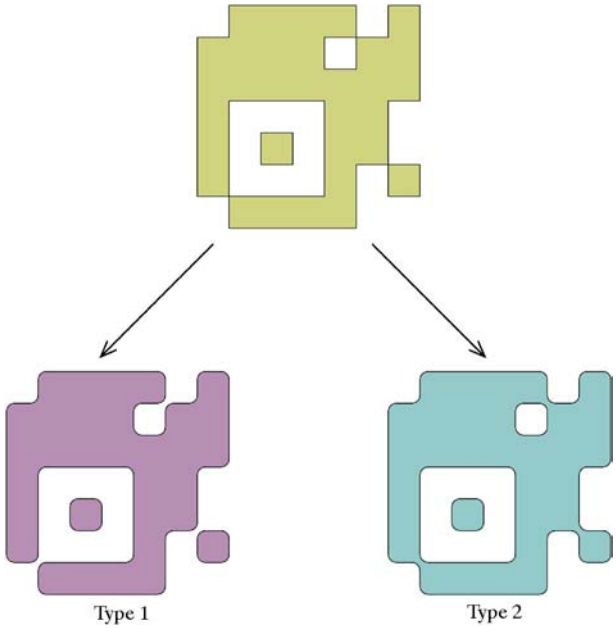

Figure 3
Removal of surface intersections. In order to calculate the Euler characteristic properly, surface intersections must be eliminated. Two alternate definitions of connectivity in a binary electron-density map are shown. An object can either be connected only if two volume elements meet face to face (type 1) or if the two volume elements touch face-to-face, edge-to-edge or vertex-to-vertex (type 2).

To measure the local distribution of electron density found in maps, three templates were chosen under which we calculated the fractional occurrence of all possible binary patterns. The first template is a cube, two voxels in length, width and height (Fig. 4*a*). The second is also a cube, but one in which the voxels are not adjacent (*i.e.* the corner voxels in a $3 \times 3 \times 3$ cube). We denote this template type as $2 \times 2 \times 2'$. This particular template allows a larger region of the map to be characterized without introducing an unreasonably large number of distinct patterns. The third is a slab three voxels in length and width and one voxel in height (Fig. 4*b*). The voxels that make up a template are considered to be either full or empty (*i.e.* either above or below the chosen contour level).

The total number of binary patterns that can be derived from a $2 \times 2 \times 2$ cube is 256. However, for the present application, patterns related by a rotation are taken to be equivalent. Though it is not strictly true for non-orthorhombic unit cells, we assume that the template is essentially cubic in shape and so conforms to cubic (octahedral) symmetry. Under this relation, there are 23 distinct patterns for a $2 \times 2 \times 2$ cube (Fig. 4*a*). For the $3 \times 3 \times 1$ slab, the 512 possible patterns reduce to 102 distinct patterns by rotational equivalence under D_4 symmetry (Fig. 4*b*). For all three of these search templates, the number of unique patterns is small enough that one can expect a statistically significant population for each unique pattern. In addition, the three templates are independent enough in shape that they might capture different information about the tendency of different patterns to appear in electron-density maps.

The frequency of occurrence of each of the unique binary patterns for the three templates is calculated in the following manner. The template is moved systematically through the map. As the template is moved through the map, the frequencies of the geometric patterns are tabulated. To normalize for unit cells of different size, we convert the frequencies to fractional values as follows. The possible patterns are grouped into subclasses, each comprised of the unique patterns that contain an equal number of bits 'on'. The fractional occurrence of each pattern is then calculated relative to the other patterns in its subclass. This analysis is performed for each of the three templates described above.

Eight additional composite parameters are also calculated in an attempt to capture any cooperative behavior between select patterns. For each of the $2 \times 2 \times 2$ templates, three parameters are calculated. Letting N_i denote the total number of observed patterns with i bits 'on', these composite parameters are: N_4/N_8, $(N_0N_8)/(N_4N_4)$ and $(N_0N_8)/(N_1N_7)$. Finally, the frequency with which a $3 \times 3 \times 3$ or $5 \times 5 \times 5$ template occurs with no bits 'on' is also calculated.

2.4. Generation of a phase-error-predicting formula

For every map generated in our simulations, the topological and geometric descriptors described in the above sections were calculated and stored. A linear function relating the

measured descriptors of a given map to a calculated phase error is defined as

$$\varphi_j^{\rm calc} = \sum_{i=1}^{172}[w_{ij}(P_{ij} - P_{ij}^o)], \qquad (7)$$

where j is one of the eight binary cutoff levels for conversion to the binary electron-density map, w_{ij} is a linear weight associated with the ith map descriptor for the jth binary cutoff level, P_{ij} is the value of the ith map descriptor at the jth binary cutoff level and P_{ij}^o is the value of the corresponding descriptor at zero phase error. In our test calculations, we have assumed that the target values P_{ij}^o are known precisely. In real applications, this may not be the case. However, uncertainties in P_{ij}^o only affect the calculated phase error by an additive constant, so relative changes in phase error are unaffected. These changes in calculated phase error are of primary interest here.

There are 172 descriptors: 16 topological, 23 for each of the $2 \times 2 \times 2$ templates with different spacings, 102 for the $3 \times 3 \times 1$ template and eight descriptors based upon the cooperative behavior of selected geometric patterns. The descriptor weights w_{ij} were calculated in the following manner. The descriptor values were tabulated from the family of maps described in the previous section. At each of the eight contour levels, there are 4050 maps in our database with varying degrees of error (18 average phase errors $\times$ 25 randomizations $\times$ nine proteins). Weights for the descriptors were derived *via* least-squares fitting the calculated phase errors from the large number of maps to the known values by minimizing (at every cutoff level) the residual errors in the set of equations

$$\sum_{i=1}^{172}[W_{ij}(P_{ij} - P_{ij}^o)] - \varphi_j^{\rm known} = 0. \qquad (8)$$

Among the descriptors tested, not all are necessarily independent. The lack of linear independence would cause the linear least-squares equations to be degenerate. To overcome any possible degeneracy, we employed eigenvalue filtering (a form of singular value decomposition) during the derivation of weights. From this filtering method, the top 50–70 components were used in the algorithm. This form of filtering removes combinations of parameters that contribute little to the phase-prediction formula.

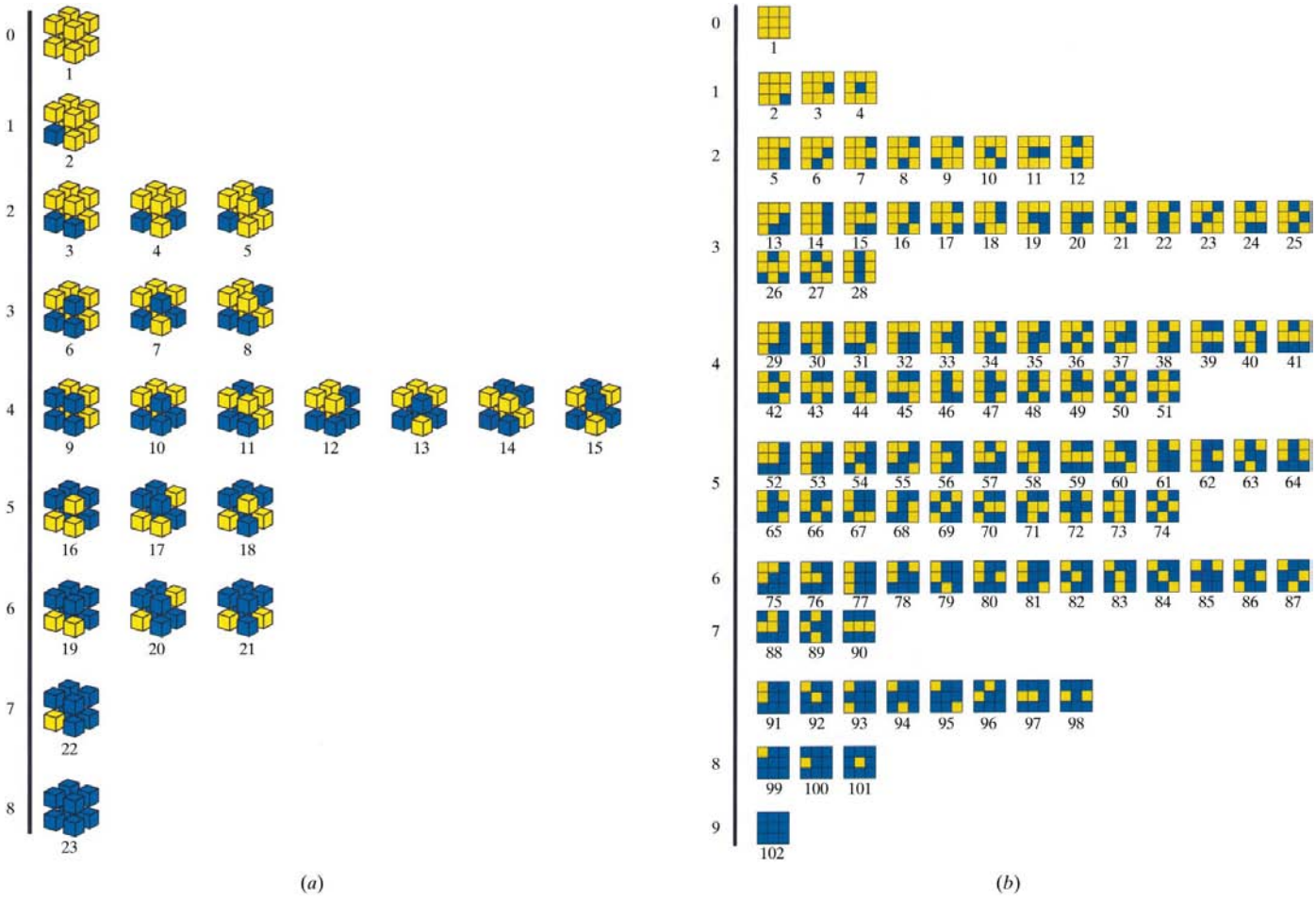

Figure 4
An enumeration of the distinct binary patterns possible on different symmetric templates. (*a*) The eight elements of a $2 \times 2 \times 2$ template allow 256 binary patterns. Under the cubic (octahedral) rotational symmetry of the template these reduce to 23 unique binary patterns, organized by the number of bits containing density. (*b*) The 102 unique binary patterns for a $3 \times 3 \times 1$ template (under D_4 symmetry) arranged by the number of bits containing density.

3. Results and discussion

3.1. Individual parameters

Some of the map properties investigated showed strong correlations with phase errors, while others did not. As a simple gauge of predictive power for each descriptor, a correlation coefficient between the descriptor value and the known phase error was calculated over the set of maps at the appropriate binary cutoff level. Fig. 5 shows a plot of two geometric descriptors *versus* phase error – one that strongly correlates with the error introduced to the phase sets and one that does not. Many of the geometric descriptors show correlations with phase error in excess of 0.9 (or less than −0.9). Those with the highest correlation coefficients (±0.90) are listed in Fig. 6.

The results of the geometrical analysis show general trends that might be expected. For example, patterns that are more connected or more compact (*e.g.* a box rather than a checkerboard) tend to be observed less frequently as phase error is increased, while the patterns that are more complex or disjointed tend to be more frequent with increasing phase error. This trend is most evident when examining pairs of patterns that differ by one bit or have the same number of bits 'on' but in a different arrangement (2 × 2 × 2: #19 *versus* #21, #7 *versus* #13, #7 *versus* #8, #6 *versus* #7; 3 × 3 × 1: #99 *versus* #100, #32 *versus* #52). Beyond some general trends, however, the complexity of Fig. 6 makes it clear that the detailed results of the geometrical analysis could not have been predicted in advance.

Among the topological descriptors, the Euler characteristic (χ), the total topological complexity (G_T) and the surface area to volume ratio capture the protein-like features in an electron-density map best (Table 2). The magnitudes of these descriptors increase with the addition of phase error at low binary cutoff levels, but decrease with increased phase error at higher cutoff levels. The opposite trend is observed for the number of discrete surfaces (nsurf) and the volume of the largest continuous region of low density. As before, while some of the observed trends can be rationalized, others cannot. A strength of the present approach is that it makes no *a priori* assumption about what patterns or properties should be favored.

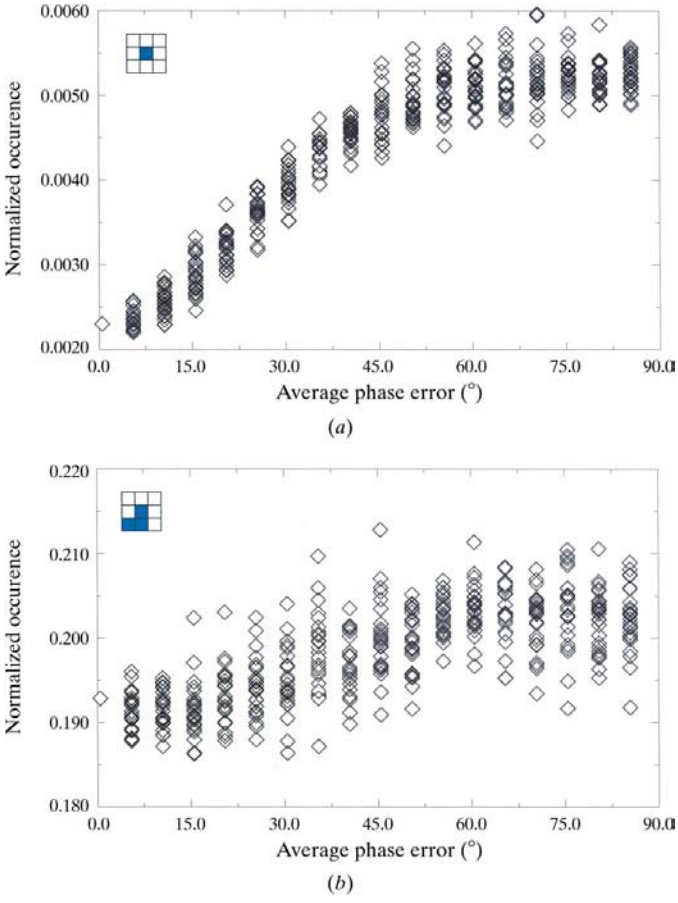

Figure 5
Examples of map descriptor values *versus* phase error. (*a*) A geometric descriptor that varies significantly with the addition of phase error to an electron-density map [the normalized occurrence of pattern #4 of the 3 × 3 × 1 template, binary cutoff level #5 (1.2σ), protein = 5cpa]. (*b*) A geometric descriptor that does not vary predictably as noise is added to the map [normalized occurrence of pattern #13 of the 3 × 3 × 1 template, binary cutoff level #1 (0.4σ), protein = 5cpa].

3.2. Analysis of the function

As described in §2 (equations 7 and 8), the descriptors that were independent of each other were combined to give a single function. The function described in the previous section was tested for self-consistency by withholding one of the structures from the database and then recalculating the descriptor weights. With these unbiased weights, the function was used to test descriptors extracted from maps of the withheld structure in order to calculate predicted errors in the phase sets. More specifically, to calculate the predicted phase error $\varphi^{\rm calc}$, the descriptors from a map of the query protein (P_{ij}) were used with weights (w_{ij}) derived from the 'jack-knifed' database, while the target values of each descriptor at zero phase error (P^{o}_{ij}) were taken from the query protein. Typically, for average phase errors spanning the entire range (0–90°) this analysis yielded 8° RMSD error in the estimation of the phase error. A plot of the calculated *versus* the observed average phase error shows that the function is roughly biphasic, with two approximately linear regions, separated by a transition near 60° (Fig. 7).

To optimize the predictive quality of the function, linear weights were calculated separately for several phase-error ranges. The results of these analyses are shown in Table 3. Subsequent analyses utilized weights calculated for the linear region between 0 and 60° average phase error, where the RMSD error in phase-error estimation is only 3.3°.

3.3. Heavy-atom searches

We tested the ability of the method (referred to hereafter as TGA) to identify correct heavy-atom positions given native diffraction amplitudes and amplitudes from an isomorphous heavy-atom derivative. For the heavy-atom searches, SIR data from a platinum derivative of the ycaC gene product from *Escherichia coli* (Colovos *et al.*, 1998) were used to generate

Table 2
Topological parameters that correlate best with change in phase error.

Parameter values increase as phase error increases, except for the parameters preceded by an asterisk. The choice of definition for a connected surface is denoted by a subscript and is fully described in the text. Topological parameters in this table have a correlation coefficient greater than 0.90 with respect to phase error.

Binary cutoff level	Topological parameters
1	$G_{T(1)}$, $\chi_{(1)}$, $G_{T(2)}$, surfmax$_{(1)}$, *solvmax$_{(1)}$, *solvmax$_{(2)}$, surfmax$_{(2)}$, $\chi_{(2)}$, SA/V
2	$G_{T(2)}$, *solvmax$_{(2)}$, surfmax$_{(2)}$, $\chi_{(2)}$, SA/V, $\chi_{(1)}$
3	SA/V, *solvmax$_{(2)}$, surfmax$_{(2)}$
4	SA/V
5	*$\chi_{(1)}$, $G_{T(1)}$, SA/V
6	*$\chi_{(1)}$, *$\chi_{(2)}$, *$G_{T(1)}$, solvmax$_{(2)}$
7	solvmax$_{(2)}$, *$\chi_{(1)}$, *$\chi_{(2)}$, *$G_{T(1)}$
8	solvmax$_{(2)}$, *SA/V, *$\chi_{(2)}$, *$\chi_{(1)}$, *$G_{T(1)}$, nsurf$_{(2)}$, npos$_{(2)}$

Table 3
Error in estimation of average phase error.

$\Delta\varphi$ is an estimation of average phase error. $\Delta\varphi_{obs}$ is based on the difference between the phases from the final refined atomic model and the phases after different stages of density modification. $\Delta\varphi_{calc}$ is the predicted phase error from map evaluation by the TGA method.

Chosen range for average phase error (°)	RMSD error in estimation of average phase error (°)
0–90	7.6 ± 1.9
0–60	3.3 ± 0.9
20–60	3.4 ± 0.9
60–90	11.3 ± 6.2

trial maps. The derivative contains two sites per asymmetric unit in space group $P42_12$ and gives very good phase information (for data to 2.0 Å phasing power = 1.88, R_{cullis} = 0.63). All calculations were performed using data to 3.0 Å resolution. Phases were calculated from the heavy-atom positions using the program *MLPHARE* (Otwinowski, 1991) and the resulting electron-density maps were examined as described in the preceding sections.

We began by using only one of the two heavy-atom sites. According to the TGA method for estimating phase error from electron-density maps, a single correct heavy-atom position had a calculated phase error that was 1.8σ less than the mean of phase errors from maps generated from randomly chosen heavy-atom sites. For this calculation, the weights were derived from the parameters of the nine database structures. Given that there are two heavy-atom sites in the asymmetric unit, it is not surprising that a map generated from a single heavy-atom site did not give a calculated phase error drastically different from the noise level.

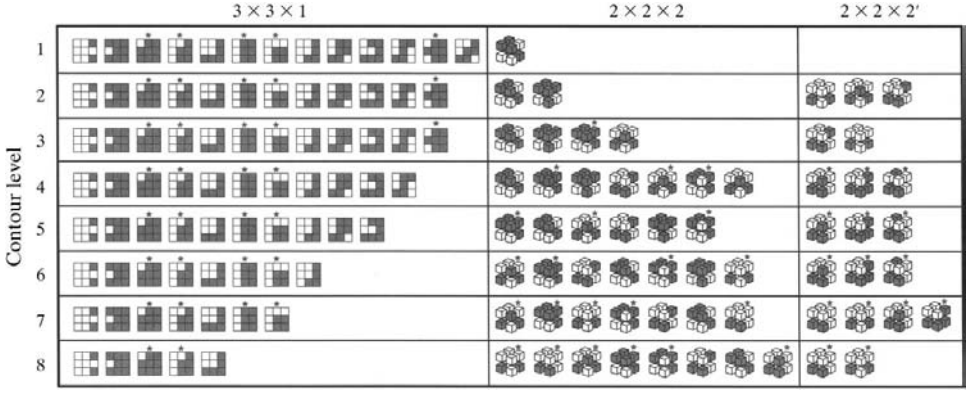

Figure 6
The geometric patterns with the strongest correlation to phase error. The table is organized by template type (3 × 3 × 1, 2 × 2 × 2 and 2 × 2 × 2′) and by binary cutoff level. The occurrence of a pattern increases as phase error is increased, except for cases denoted with an asterisk, where the trend is reversed. Geometric parameters listed in this figure have a correlation coefficient greater than 0.90 with respect to phase error.

Further calculations were performed using two sites where one of the two heavy atoms was fixed at its correct position while the other was allowed to vary. In this case, the correct heavy-atom position gave a map whose calculated phase error was 5.7σ less than the mean calculated phase errors of maps generated from phases where both sites were randomly positioned (Fig. 8). The performance of TGA with heavy-atom data suggests that it may be useful in evaluating potential heavy-atom models.

In order to compare the discriminatory power of TGA with more traditional ideas from density modification, the heavy-atom calculations above were repeated and the same maps were evaluated with a histogram-matching protocol. We calculated a correlation coefficient between the electron-density histogram of each candidate map and the histogram for an ideal map (Zhang & Main, 1990). Following the same reasoning as with TGA, if a map derives from accurate phases, then its density histogram should be highly correlated with the ideal histogram. With the same two-site heavy-atom model as before, the histogram correlation is 5.1σ higher than the average for maps based on two random sites. The apparently comparable power of TGA and histogram matching suggests that TGA might be useful as a density-modification tool if it can be implemented as an iterative procedure or combined with existing density-modification procedures.

3.4. Evaluating the progress of density modification

A second test was performed to assess whether or not the TGA function could accurately discriminate between phase sets arising from different methods of density modification. The platinum SIRAS data (using both sites) of the ycaC gene product was subjected to either solvent flattening, histogram matching, solvent flattening and histogram matching, symmetry-averaging and histogram matching, or solvent flattening, averaging and histogram matching, with the

Table 4
Comparison of observed and calculated phase errors after various density-modification procedures.

Phase set	$\Delta\varphi_{obs}$ (°)	$\Delta\varphi_{calc}$ (°)
SIRAS	44.1	46.1
SIRAS plus solvent flattening alone	40.0	37.3
SIRAS plus histogram matching alone	40.9	29.3
SIRAS plus solvent flattening and histogram matching	39.9	24.3
SIRAS plus NCS averaging and histogram matching	34.2	16.1
SIRAS plus solvent flattening, NCS averaging and histogram matching	31.2	12.2

program *DM* (Cowtan, 1994). The resulting electron-density maps were subjected to the estimated phase-error calculation. For comparison, an estimate of the true error in each phase set was obtained by assuming that the phases calculated from the final atomic structure are essentially correct (Table 4). In all cases, the calculated phase error went down, as a result of density modification. Furthermore, the TGA function was able to correctly predict the relative accuracy of the various phase sets produced by different density-modification schemes.

On closer inspection, it is evident that the TGA function indicates a greater improvement in phase accuracy during density modification than is calculated by comparing with final model phases. In fact, this exaggerated drop may reflect difficulties pointed out earlier in establishing the correct (zero-error) map descriptors for a perfect map. On the other hand, the phase improvement calculated by the TGA method may not be far off from the correct value. It should be kept in mind that the model phases may be significantly different from the true (unknown) phases. This residual error could account for much of the discrepancy between the two estimates of final phase error. In any case, it seems clear that the TGA method could provide valuable feedback during density-modification procedures. For instance, during symmetry averaging one might monitor the drop in calculated phase error to help decide whether the non-crystallography symmetry operators have been defined accurately.

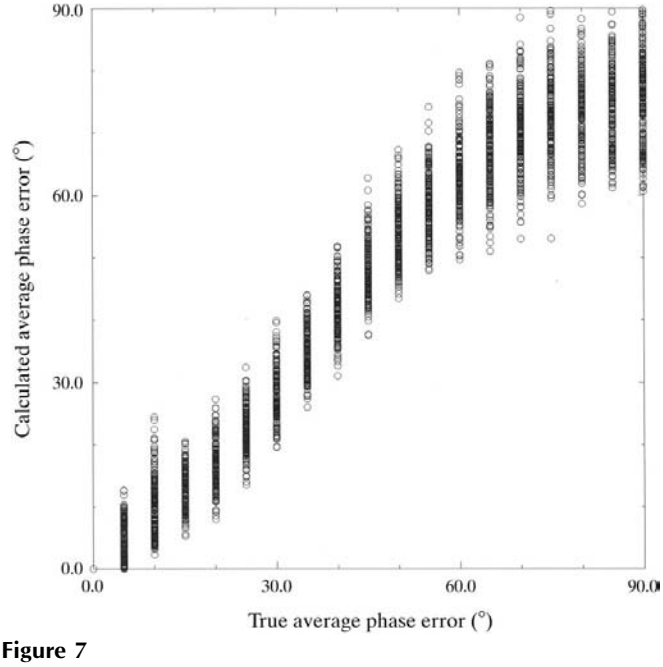

Figure 7
Plot of known *versus* calculated average phase error using an empirically derived map-evaluation function. The calculated phase errors for each map are based on weights derived from a jack-knifed database in order to avoid bias.

3.5. Conclusions

The protein-like appearance of electron-density maps can be assessed by a function based on a topological and geometrical analysis (TGA). The present method was developed from a systematic search for properties of electron-density maps that correlate with phase error. A subset of the topological and geometrical map descriptors examined were found to be informative and were used to build an empirical function that measures the accuracy of diffraction phase sets. The results demonstrate that it is possible to automatically evaluate electron-density maps and accurately distinguish between phase sets of varying quality.

In the present implementation, it was possible to determine the relative quality of maps created from experimental data and various heavy-atom substructure models. Even better results might be obtained if the function was optimized to discriminate between correct and incorrect heavy-atom

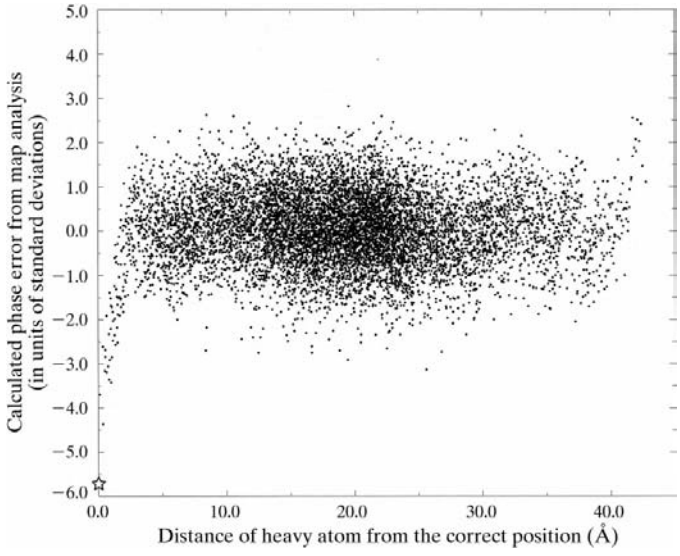

Figure 8
Plot of relative phase error (predicted with the map-evaluation function) *versus* distance of a heavy atom from its correct position. One heavy atom was placed at the correct position and another was systematically moved about the plane that contains the correct position for the second site. For each heavy-atom configuration, an electron-density map was calculated and the empirical function was used to estimate the phase error. The plot shows the number of standard deviations from the mean calculated phase error (over all maps) *versus* the Euclidean distance of the second atom from the correct site. The correct site gives a calculated phase error that is 5.7 standard deviations below the mean and is denoted by a star in the plot.

substructure solutions, as typically provided by automated Patterson searching routines. Additionally, given that even exceptional MIR maps do not possess ideal protein-like electron-density characteristics without density modification, the heavy-atom searching routine could be further optimized if it were trained solely on experimentally derived maps. Keeping this in mind, it is not surprising that the function performed unusually well when tested on various density-modified maps. Since density-modification techniques are designed to make electron density appear more ideal, the effect of density modification on the behavior of the function seems to be even more marked than the effect of altering the heavy-atom substructure.

The TGA method of evaluation could be useful in numerous applications, such as validating non-crystallographic symmetry operators in symmetry-averaging and automating HA substructure solution. The estimates of phase error could also help to provide more accurate figures of merit than those typically returned with current density-modification procedures. Finally, although it has not been so tested in this work, the function might serve as a target function for density modification or (at lower resolutions) as a tool for *ab initio* phasing of macromolecular structures. A web-based version of the TGA algorithm that compares the quality of two maps is available at http://www.doe-mbi.ucla.edu/people/colovos/TGA.

We thank Drs Robert Grothe and Edward Marcotte for useful comments and discussion. This work was supported by NSF grant NSF-DBI 98-07896 to TOY, NIH training grant GM07185 (EAT) and NIH training grant GM07185 (CC).

References

Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–439.
Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Colovos, C., Cascio, D. & Yeates, T. O. (1998). *Structure*, **6**, 1329–1337.
Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newslett. Protein Crystallogr.* **31**, 34–38.
Goldstein, A. & Zhang, K. Y. J. (1998). *Acta Cryst.* D**54**, 1230–1244.
Hauptman, H. (1986). *Science*, **233**, 178–183.
Ioerger, T. R., Holton, T. R., Christopher, J. A. & Sacchettini, J. C. (1999). *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, edited by T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes & R. Zimmer, pp. 130–137. Menlo Park, CA: AAAI Press.
Karle, J. (1986). *Science*, **232**, 837–843.
Kleywegt, G. L. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.
Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.
Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (1998). *Acta Cryst.* D**54**, 1078–1084.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 1872–1877.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst.* A**46**, 41–46.