Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 21 July 2000
Accepted 13 October 2000

An exponential modeling algorithm for protein structure completion by X-ray crystallography

V. L. Shneerson,[a] D. L. Wild[b] and D. K. Saldin[a]*

[a]Department of Physics, University of Wisconsin-Milwaukee, PO Box 413, Milwaukee, WI 53201, USA, and [b]Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont, CA 91711, USA. Correspondence e-mail: dksaldin@csd.uwm.edu

An exponential modeling algorithm is developed for protein structure completion by X-ray crystallography and tested on experimental data from a 59-residue protein. An initial noisy difference Fourier map of missing residues of up to half of the protein is transformed by the algorithm into one that allows easy identification of the continuous tube of electron density associated with that polypeptide chain. The method incorporates the paradigm of phase hypothesis generation and cross validation within an automated scheme.

1. Introduction

Determination of the phases of measured Bragg reflections is the fundamental theoretical problem of X-ray crystallography. In protein crystallography, the three most commonly used techniques for obtaining an initial estimate of these phases are (*a*) the isomorphous replacement method (Green *et al.*, 1954; Blow & Crick, 1959), which requires the preparation of related structures in which heavy atoms are attached to the original protein; (*b*) the multiple-wavelength anomalous-dispersion (MAD) method (Hendrickson, 1991; Leahy *et al.*, 1992), which depends on the presence of sufficiently strong anomalously scattering atoms within the protein; and (*c*) the molecular replacement method (Rossmann & Blow, 1962), which requires the identification of a *known* structure similar to the one whose structure is being sought. Initial phases obtained by any of these methods must be refined for a successful structure determination. When an interpretable electron-density map is obtained, an initial atomic model may be built and refined against the observed data.

During this procedure, the necessity for *structure completion* may arise in several ways: for instance, a partial model of the protein molecule may have been constructed during model building and refinement, or molecular replacement may have been carried out with a 'probe' similar to only a fragment of the target molecule. In the latter case, suppose that a molecule or molecular fragment of known structure may be identified similar to the unknown one to be determined. The first step in using this information to solve the unknown structure is to perform a rotation and translation search to orient the known 'probe' to match that of its counterparts in the structure to be determined. The next step is the recovery of the missing part of the unknown structure and the refinement of the phases. If the known part is a substantial portion of the total, the *difference Fourier* method (Cochran, 1951) enables the missing part of the structure to be determined with reasonable accuracy.

In the case of *structure refinement*, *omit maps* based on the difference Fourier synthesis may be calculated from partial structures obtained by the omission of parts of the model suspected of being in error (see *e.g.* Drenth, 1994). Read (1986, 1997) has described weights for the structure-factor amplitudes for Fourier syntheses which reduce model bias in the case of an incomplete or partially incorrect structure and Hodel *et al.* (1992) have shown that model bias in omit maps may be reduced by the refinement of the partial structure.

Another method of automated structure completion and refinement, comparable to the iterative least-squares minimization/difference Fourier synthesis approach used in small-molecule crystallography, has been proposed by Lamzin & Wilson (1993). This procedure requires data of high quality and a partial structure of approximately 75% of the total molecule. A variation of this approach has also been used by Fitzgerald (1994) to automatically fit non-peptide (ligand) electron density in protein–ligand complexes.

An alternative idea for structure completion that has been proposed recently (Szöke, 1993; Maalouf *et al.*, 1993; Somoza *et al.*, 1995) exploits the analogy with *holography* (Gabor, 1948; Collier *et al.*, 1971). The X-ray diffraction pattern is likened to a hologram formed by interfering a known *reference wave* from the known part of the structure with an unknown *object wave* from the unknown part of the structure. The aim of holographic reconstruction is to determine the object wave from a knowledge of the reference wave and the experimental data. From a full knowledge of the object wave, the electron density of the unknown part of the structure may be determined by Fourier transformation. Szöke and co-workers have proposed an algorithm for determining that unknown electron density, provided the X-ray diffraction data are supplemented by some other constraints, chief of which are the positivity of the electron density and some prior knowledge of the molecular envelope (or conversely the solvent regions).

In practical applications of their method, Maalouf *et al.* (1993) and Somoza *et al.* (1995) have used a conjugate-gradient algorithm to minimize an appropriately defined *cost function* quantifying the fit of a theoretical model to the data subject to the constraints. Any kind of gradient search algorithm requires the starting point of the (presumably multipeaked) cost function to lie within a basin of convergence of the global minimum in the multidimensional search space. Limitations of the data set, the sudden cut-off of the data at the outer limits of experimentally accessible Bragg reflections, other missing reflections, and experimental errors can give rise to ill-conditioned solutions for the unknown electron density. Indeed, Szöke (1993) has argued that the ill-conditioning problem in X-ray crystallography is intrinsic and related also to the fact that the data are available only at discrete points in reciprocal space (the Bragg points). In order to overcome some of these problems, Szöke proposed representing the electron density as a linear superposition of Gaussian functions at each of a set of grid points. This introduces an extra set of parameters associated with the choice of the width of the Gaussians. In their realistic first tests of their algorithm to recover protein fragments forming a substantial fraction of the whole unit cell, Somoza *et al.* (1995) took as their cost function to be minimized at each iteration a combination of a linearized holographic discrepancy function ($f_{\rm eden}$), a function ($f_{\rm space}$) that minimizes the discrepancy to a 'target' density taken from a smeared-out version of a model of the protein, solved previously by standard crystallographic methods, and another function ($f_{\rm null}$) that minimizes the projection of the recovered electron density onto 'the null space of the encoding operator'. The three component cost functions are then summed in the ratio of further parameters, termed Lagrange multipliers. Each of the three component cost functions contained further parameters (termed 'weights') whose values were estimated by other complicated arguments, or else by 'strengths of belief'. Although in a later paper, through a test on synthetic data for the thaumatin molecule, Szöke *et al.* (1997) suggested that use of a quadratic discrepancy function (Saldin *et al.*, 1993) may eliminate the need for the $f_{\rm null}$ cost function, even this algorithm requires the estimation of several parameters, such as a Lagrange multiplier and those defining the Gaussian basis functions.

In the present paper, we develop an alternative method for structure completion inspired by the principles of Bayesian statistics (see *e.g.* Sivia, 1996) which provide a prescription for making optimal objective inferences from limited data and other prior knowledge. The method combines features of Szöke's holographic method with those of another technique that has made a major mark in modern crystallographic phasing, namely the maximum-entropy method (Jaynes, 1957).

In particular, we adapt the exponential modeling algorithm proposed by Collins (1982) to address the structure completion problem. We show that this enables high-quality structure completion from a knowledge of as little as 50% of the total structure. As such, the method is most likely to be of use for molecular replacement. A preliminary report of our method has been published previously (Saldin *et al.*, 1997), where we describe an application to the problem of structure completion from synthetic (*i.e.* calculated) diffraction data for the protein bovine pancreatic trypsin inhibitor (BPTI) (also considered by Maalouf *et al.*, 1993). In the present paper, we apply our scheme to previously published experimental data for a 59-residue protein.

For the structure completions we have attempted, it was not possible to clearly identify the chain of missing residues from either a Sim-weighted difference Fourier map (Sim, 1959, 1960) or a difference Fourier map calculated with phases and weights of the entire protein from a standard *density modification* program (Cowtan & Main, 1998) that starts from such experimental phases as from multiple isomorphous replacement (MIR). Nevertheless, our exponential modeling algorithm was able to improve either of these starting maps to the point where the outline of this part of the polypeptide chain was found to stand out distinctly from the surrounding background noise.

A similar algorithm has been developed for the analogous problem of finding the electron density of a surface from a knowledge of the bulk structure in surface X-ray diffraction (SXRD) (Saldin *et al.*, 2000).

We begin in §2 by a mathematical statement of the structure completion problem. In §3, we describe several different difference Fourier methods for structure completion. §4 describes applications of difference Fourier methods to structure completion from experimental data from crystallized α-dendrotoxin from green mamba (*Dendroapsis angusticeps*) venom (Protein Data Bank entry 1DTX) (Skarzynski, 1992). We develop our exponential modeling algorithm in §5, and §6 describes its application to the same experimental data. §7 contains a discussion and §8 our conclusions.

2. The structure completion problem

Suppose that the unit cell of a crystal is divided into a set of voxels centered on a uniform grid of points. Let the number of electrons from the known part of the structure in the voxel centered on the position $\mathbf{r}_i$ be n_i. Then the contribution from the known part of the structure (also termed the partial structure) to the structure factor of the Bragg reflection $\mathbf{g}$ will be given by the discrete Fourier transform

$$R_{\mathbf{g}} = \sum_i n_i \exp{(i\mathbf{g}\cdot\mathbf{r}_i)}. \qquad (1)$$

The quantities $R_{\mathbf{g}}$ may be regarded as the Fourier coefficients of a holographic *reference* wave. If the corresponding Fourier coefficients from the unknown part of the electron distribution in the unit cell are represented by $O_{\mathbf{g}}$ (which may be regarded as the coefficients of an *object* wave), the total intensity of the Bragg reflection may be written

$$I_{\mathbf{g}} = |F_{\mathbf{g}}|^2, \qquad (2)$$

where the structure factor $F_{\mathbf{g}}$ may be written as the sum

$$F_{\mathbf{g}} = R_{\mathbf{g}} + O_{\mathbf{g}}. \qquad (3)$$

The recovery of the unknown coefficients $O_{\mathbf{g}}$ from a set of measured intensities $I_{\mathbf{g}}$ and the known coefficients $R_{\mathbf{g}}$ is the classic problem of holography. It is this analogy that has been explored by Szöke and co-workers. Of course, once the complete set of object wave components $O_{\mathbf{g}}$ is recovered, the electron distribution $\{u_i\}$, defined on the same voxel grid, may be found by an inverse Fourier transform.

If both the amplitudes *and* phases of the structure factors $\{F_{\mathbf{g}}\}$ are known, the unknown electron distribution may be recovered directly from the formula

$$u_i = (1/N)\sum_{\mathbf{g}}\{F_{\mathbf{g}} - R_{\mathbf{g}}\}\exp(-i\mathbf{g}\cdot\mathbf{r}_i), \tag{4}$$

where N is the number of voxels per unit cell. The spatial resolution of the electron distribution is determined by the maximum magnitude of the reciprocal-lattice vector of the non-zero structure factors in the sum (4). The problem, of course, is that although the amplitudes of $\{F_{\mathbf{g}}\}$ are directly measurable from the experimental data their phases are not. The class of techniques termed difference Fourier methods which have been developed to address this problem are discussed next.

3. Difference Fourier syntheses

3.1. Unweighted

If the unknown part of the structure is not too large a proportion of the whole, a reasonable estimate of it may be obtained by the *unweighted difference Fourier* method (Cochran, 1951), which approximates the phases of the structure factors by those of the known part of the structure, *i.e.* it estimates the electron distribution of the unknown part by

$$u_i^{(\mathrm{UDF})} = (1/N)\sum_{\mathbf{g}}\{|F_{\mathbf{g}}|\exp(i\varphi_{\mathbf{g}}^{(R)}) - R_{\mathbf{g}}\}\exp(-i\mathbf{g}\cdot\mathbf{r}_i), \tag{5}$$

where

$$\varphi_{\mathbf{g}}^{(R)} = \arg(R_{\mathbf{g}}), \tag{6}$$

the phase of $R_{\mathbf{g}}$, which *is* known since it is derived from a calculation of $R_{\mathbf{g}}$ from the known part of the structure.

3.2. Weighted

As pointed out by Read (1986), in general a better estimate of the electron distribution may be obtained from the expression

$$u_i^{(\sigma_A)} = (1/N)\sum_{\mathbf{g}}\{m_{\mathbf{g}}|F_{\mathbf{g}}|\exp(i\varphi_{\mathbf{g}}^{(R)}) - D_{\mathbf{g}}R_{\mathbf{g}}\}\exp(-i\mathbf{g}\cdot\mathbf{r}_i), \tag{7}$$

where $m_{\mathbf{g}}$ is a *figure of merit* that represents the average effect of possible deviations of the phase of $F_{\mathbf{g}}$ from $\varphi_{\mathbf{g}}^{R}$ and $D_{\mathbf{g}}$ takes account of all possible sources of uncertainty in the coordinates of the partial structure. The *SIGMAA* computer program (Read, 1986) calculates both of these quantities from a set of structure factors $\{R_{\mathbf{g}}\}$ of the partial structure and experimental amplitudes $|F_{\mathbf{g}}|$ of the entire structure.

In the limit where the partial structure may be considered perfectly known, $D_{\mathbf{g}} = 1$ and $m_{\mathbf{g}} = w_{\mathbf{g}}^{(\mathrm{Sim})}$, a weighting factor previously derived by Sim (1959, 1960). Then Read's formula (7) reduces to the *Sim-weighted difference Fourier* (SWDF) expression:

$$u_i^{(\mathrm{SWDF})} = (1/N)\sum_{\mathbf{g}}\{w_{\mathbf{g}}^{(\mathrm{Sim})}|F_{\mathbf{g}}|\exp(i\varphi_{\mathbf{g}}^{(R)}) - R_{\mathbf{g}}\}\exp(-i\mathbf{g}\cdot\mathbf{r}_i), \tag{8}$$

where

$$w_{\mathbf{g}}^{(\mathrm{Sim})} = I_1(X)/I_0(X) \tag{9}$$

for non-centric reflections,

$$w_{\mathbf{g}}^{(\mathrm{Sim})} = \tanh(X/2) \tag{10}$$

for centric ones, and

$$X = 2|F_{\mathbf{g}}||R_{\mathbf{g}}|\Big/\sum_{i=1}^{n} f_i^2. \tag{11}$$

In the expressions above, $I_0(X)$ and $I_1(X)$ are modified Bessel functions of order zero and one, and the f_is are the scattering factors of the missing atoms.

In all of the above formulae (5), (7) and (8), the phase assigned to the structure factor $F_{\mathbf{g}}$ of the whole protein is that of the partial structure. An alternative approach is to assign to the phases of the entire protein values determined by a procedure termed density modification. This will be discussed next.

3.3. Density-modified difference Fourier

Measured structure factors from protein crystals generally contain contributions from a disordered water solvent, in addition to those from the protein molecule. An established method of refining initial phases in protein crystallography (obtained from *e.g.* MIR or MAD experiments) is the combination of solvent-flattening and histogram-matching techniques, known as the density-modification (DM) method (Cowtan & Main, 1998). This procedure may be conveniently implemented by *e.g.* the CCP4 routine *dm* (Collaborative Computational Project, Number 4, 1994). The output of this routine includes an improved set of phases $\{\varphi_{\mathbf{g}}^{(\mathrm{DM})}\}$ and an associated set of weighting coefficients, $w_{\mathbf{g}}^{(\mathrm{DM})}$, for the structure factors of the entire protein from which one may calculate an improved estimate of the electron distribution $\{u_i^{\mathrm{DM}}\}$ of the entire contents of the unit cell (including solvent) by

$$u_i^{\mathrm{DM}} = (1/N)\sum_{\mathbf{g}} w_{\mathbf{g}}^{(\mathrm{DM})}|F_{\mathbf{g}}|\exp(i\varphi_{\mathbf{g}}^{(\mathrm{DM})})\exp(-i\mathbf{g}\cdot\mathbf{r}_i). \tag{12}$$

This suggests the alternative form of difference Fourier synthesis which we may term a *density-modified difference Fourier* (DMDF), defined by

$$u_i^{(\mathrm{DMDF})} = (1/N)\sum_{\mathbf{g}}\{w_{\mathbf{g}}^{(\mathrm{DM})}|F_{\mathbf{g}}|\exp(i\varphi_{\mathbf{g}}^{(\mathrm{DM})}) - R_{\mathbf{g}}\}\exp(-i\mathbf{g}\cdot\mathbf{r}_i), \tag{13}$$

which subtracts the electron distribution of the known part from an estimate of that of the entire unit cell.

The above and all succeeding formulae that involve a difference between an experimental structure factor, $F_{\mathbf{g}}$, and a calculated one, $R_{\bf g}$, do of course require that the experimental structure factors are placed on the same absolute scale as the theoretical ones. This may be accomplished by the well established techniques of Wilson scaling (see *e.g.* Drenth, 1994).

In the next section, we attempt to recover the electron distribution of deleted residues of the protein α-dendrotoxin using a knowledge of just parts of that structure and the published experimental data, using both the SWDF, appropriate for the case of a model (effectively error-free) partial structure, which assigns the phases of the partial structure to the structure factors of the entire protein, and the DMDF, which estimates the corresponding phases (and their associated weights) from a density-modification program that starts with experimental phases. The results will be compared to the known electron distribution of those residues. This will illustrate the limitations of difference Fourier methods when an attempt is being made to recover a large portion of the protein as in the molecular replacement method.

4. Structure completion of α-dendrotoxin by difference Fourier syntheses

α-dendrotoxin from green mamba venom (Protein Data Bank entry 1DTX) is a 59-residue protein which crystallizes into $P2_12_12_1$ space-group symmetry with unit-cell parameters $a = 73.58$, $b = 38.73$ and $c = 23.19$ Å. Skarzynski (1992) has collected experimental data to approximately 2.5 Å resolution with the maximum Miller indices of $|h_{\rm max}| = 31$, $|k_{\rm max}| = 16$ and $|l_{\rm max}| = 10$. Structure-factor amplitudes and MIR phases as distributed with the CCP4 package (Collaborative Computational Project, Number 4, 1994) were the input to the calculations reported in this paper. From the given data in the positive octant of reciprocal space, application of the appropriate symmetry relations extended the data to all eight octants.

All Fourier transforms of the present work were performed with the fast-Fourier-transform package *CFFT99* developed for the CRAY supercomputer, which requires the length of the transforms to be powers of 2, 3 or 5. For this reason, we expanded our reciprocal-space arrays to dimensions of $64 \times 32 \times 24$, covering all eight octants and initializing to zero all the structure factors with unassigned values. The corresponding real-space grid spacings in the directions of the unit vectors a, b and c were 1.15, 1.21 and 0.97 Å, respectively.

For α-dendrotoxin, which is a solved structure (Skarzynski, 1992), one can estimate the accuracy of this procedure for recovering the electron distribution of the entire protein by computing a correlation coefficient C between a DM map $\{u_i^{\rm DM}\}$ calculated from (12) and the electron distribution calculated from the atomic model in the Protein Data Bank (which we shall henceforth term the 'exact' map). We employed a linear correlation coefficient C (Press *et al.*, 1992) as a statistical measure of the agreement between the two maps. For two distributions $\{u_i\}$ and $\{v_i\}$ $(i = 1, 2, \ldots, N)$, sampled in real space, C is just the cosine of the angle between two N-dimensional vectors $\mathbf{U}$ and $\mathbf{V}$ with components $u_i - \langle u \rangle$ and $v_i - \langle v \rangle$, respectively, *i.e.*

$$C = \mathbf{U} \cdot \mathbf{V} / |\mathbf{U}||\mathbf{V}|. \qquad (14)$$

The relatively high resulting value, $C = 0.65$, is a confirmation of the value of the density modification method in constructing a first electron distribution map of the molecule.

We tested this by taking $R_{\bf g}$ to be the Fourier coefficients of the diffracted wave from successively smaller fragments of the molecule by deleting from the known structure model an increasing number of residues, starting from residue 1. Then we reconstructed the electron distribution $\{u_i^{\rm (DMDF)}\}$ from (13) for each of these sets of deleted residues. The results are shown in Fig. 1, where the line labelled DMDF$^{\rm (exp)}$ shows the resulting correlation coefficient C, between $\{u_i^{\rm (DMDF)}\}$ and the 'exact' electron distribution of the deleted residues from the known structure.

Also shown in Fig. 1 is the line labelled SWDF$^{\rm (exp)}$, which plots the correlation, C, between the reconstruction by the SWDF formula (8) of the same reconstructed fragments and those of the corresponding 'exact' densities.

One feature of these results is in marked contrast to our earlier results on structure completion with synthetic data for BPTI (Saldin *et al.*, 1997). With synthetic data, the correlation coefficient is highest for smaller recovered fragments. This is easily understood since in this case the only uncertainties are the phases of the total structure factors and since the starting guesses of the phases are those of the known part of the structure. Obviously these guesses are best for larger known parts, *i.e.* where the fragment to be recovered by structure completion is small. Indeed, in our earlier work, we found

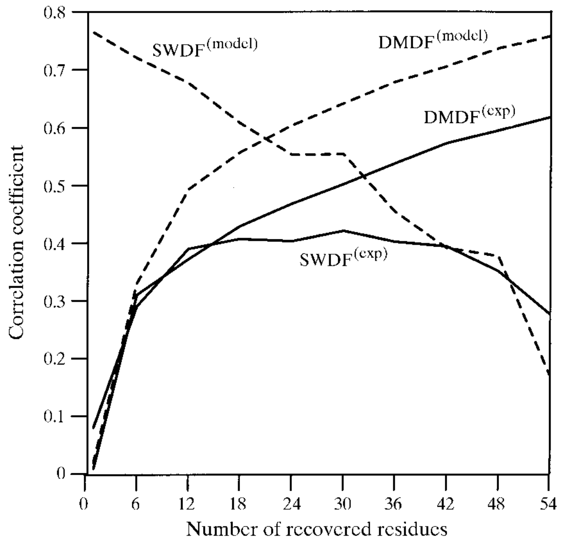

Figure 1
Linear correlation coefficients comparing the 'exact' electron distribution of various sets of missing residues with the distribution recovered by the Sim-weighted difference Fourier (SWDF) and the density-modified difference Fourier (DMDF) methods. The superscript (exp) indicates that the calculation used experimental diffraction amplitudes, while (model) indicates that the amplitudes used were those calculated from Skarzynski's model of the α-dendrotoxin molecule from the green mamba venom (Protein Data Bank entry 1DTX).

higher values of C for smaller recovered fragments. The line DMDF$^{(\rm exp)}$ of Fig. 1 displays exactly the opposite trend, with higher C's for larger recovered fragments. The line SWDF$^{(\rm exp)}$ also shows C increasing from smaller through larger recovered fragments, although in this case it seems to reach a maximum for the recovery of 30 residues and declines when an attempt is made to reconstruct a greater number. Indeed, in the limit of the recovery of just residue 1, C was found to be only 0.08 with the use of the DMDF equation (13), and just 0.01 with the Sim formula (8)!

A little reflection makes clear the reason for this difference between the results from synthetic and experimental data. In the case of the latter, there is one extra source of potential uncertainty, namely experimental errors in measurements of the structure-factor amplitudes $|F_{\bf g}|$. A glance at either the DMDF expression (13) or the SWDF formula (8) reveals that the *fractional* error in the recovered distributions, $\{u_i^{(\rm DMDF)}\}$ or $\{u_i^{(\rm SWDF)}\}$, will increase for smaller fragments recovered since the magnitudes of the *differences* of the structure factors on the right-hand sides (RHS's) of these equations are smallest in such cases.

It is easy to confirm this supposition by repeating the calculations of DMDF$^{(\rm exp)}$ and SWDF$^{(\rm exp)}$, but with the 'exact' structure-factor amplitudes from the known model of the entire protein in place of the experimental values, $|F_{\bf g}|$, in (13) and (7), respectively. The results are also shown in Fig. 1, where the line labeled DMDF$^{(\rm model)}$ represents the corresponding DMDF calculation, and the one labeled SWDF$^{(\rm model)}$ the corresponding SWDF calculation. As expected, the correlation coefficients of DMDF$^{(\rm model)}$, in which the 'exact' amplitudes are used, are consistently higher than those of DMDF$^{(\rm exp)}$, which uses experimental amplitudes. Likewise, the line SWDF$^{(\rm model)}$ is consistently higher than SWDF$^{(\rm exp)}$.

More interesting is the observation that the expected trend of higher correlations from smaller recovered fragments obtained in our model calculations for the synthetic BPTI data is reproduced here in the SWDF calculations represented by the line SWDF$^{(\rm model)}$. This indicates that the only causes of the reversal of the naïvely expected trend in SWDF$^{(\rm exp)}$ are the experimental errors in the measured Bragg amplitudes.

However, the trend of increasing C with the number of recovered residues in DMDF$^{(\rm exp)}$ is not reversed in the line DMDF$^{(\rm model)}$ from DMDF structure completion. The reason is that although the structure-factor amplitudes $|F_{\bf g}|$ may now be exact for DMDF$^{(\rm model)}$, unlike the SWDF case, experimental errors are still present in the RHS of (13) *via* the density-modified phases $\varphi_{\bf g}^{\rm DM}$. This still results in a greater *percentage* error in the smaller recovered fragments.

Fig. 2 illustrates the 'exact' electron distribution map of residues 1–18 of α-dendrotoxin from the atomic coordinates of this structure in the Protein Data Bank. Also shown is a ball-and-stick representation of these residues. The extra electrons in this figure not enclosing the ball-and-stick model are due to other symmetry-related portions of the same residues in the unit cell (which contains four molecules of the protein).

Fig. 3 shows the electron distribution of these residues as reconstructed from the SWDF formula (8). Its degree of agreement with the 'exact' electron distribution of the same residues is characterized by a value of $C = 0.41$. Fig. 4 illustrates the corresponding distribution recovered by the DMDF formula (13), whose agreement with the 'exact' map is characterized by $C = 0.43$. Considerable noise is present on both of these maps making the identification of the molecular envelope of the missing residues difficult.

In the next section, we develop an exponential modeling algorithm, which, starting from either of these difference Fourier maps, is able to improve these estimates of the electron distributions to such an extent that they enable a fairly unambiguous identification of the envelope of the missing residues.

5. Structure completion by exponential modeling

The problem of obtaining stable and meaningful solutions from incomplete and noisy data has been addressed in a variety of fields by means of the principles of Bayesian statistics (Sivia, 1996) and the maximum-entropy method in particular (Jaynes, 1957; Gull & Daniell, 1978). In X-ray crystallography, this idea has been used to develop an *exponential modeling* algorithm (Collins, 1982; Collins & Mahar, 1983) for improving the resolution of a pre-existing electron density map of a protein. A similar exponential modeling scheme is used by Bricogne (1984, 1988, 1991) and Gilmore

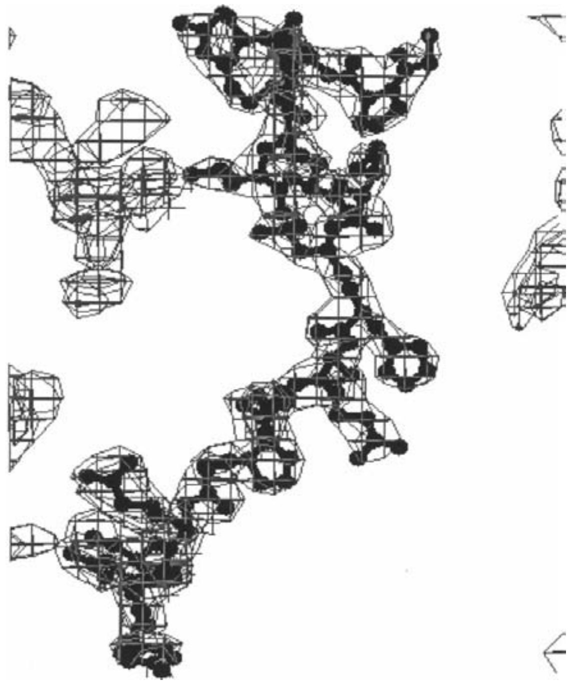

Figure 2
A view of an isosurface of an unclipped three-dimensional electron-density map of residues 1–18 of α-dendrotoxin calculated from a Fourier transform of the experimental amplitudes and the 'exact' phases calculated from the atomic model of α-dendrotoxin (Skarzynski, 1992). The wire-mesh surface corresponds to an electron density of 2.9 times the standard deviation above the mean. The isosurfaces not enclosing the ball-and-stick figure represent reconstructed electron density from parts of symmetry-related molecules in the crystal. It should be noted that these are also reproduced in the exponential modeling map of Fig. 6.

(1996) as part of an iterative process of *phase extension* in which a knowledge of the phases of some low-resolution structure factors is extended to those of higher-resolution shells as implemented by the *BUSTER* computer program (Bricogne, 1993). The potential use of maximum-entropy reconstruction has also been discussed in a recent paper by Szöke (1998). Carter & Xiang (1997) have reviewed an algorithm they term maximum-entropy solvent flattening (MESF) that incorporates solvent flattening into a maximum-entropy phase-extension algorithm.

In this paper, we adapt Collins's (1982) exponential modeling algorithm to the protein structure completion problem and demonstrate its efficacy by an application to published experimental data. The input to the algorithm consists only of the experimental structure-factor amplitudes and an initial estimate of phases and weights of the reflections. These may be obtained, for example, from multiple isomorphous replacement followed by the standard density modification procedure that combines solvent flattening with electron-density histogram matching (Cowtan & Main, 1998). Our program combines this initial phasing of the entire molecule with information on as little as half of the structure to recover the remaining molecular electron distribution to high accuracy. In a sense, the procedure may be thought of as a combination of the holographic ideas of Szöke with Collins's exponential modeling approach. This is to be contrasted with the scheme of Xiang *et al.* (1993), which combines solvent flattening with Bricogne's (1993) Bayesian scheme for progressively phasing reflections of higher resolution by a combination of exponential modeling and cross validation by means of a *global log-likelihood gain*. Our method combines a somewhat different exponential modeling scheme with cross validation by the use of Brünger's (1992, 1993, 1997) *free R factor*.

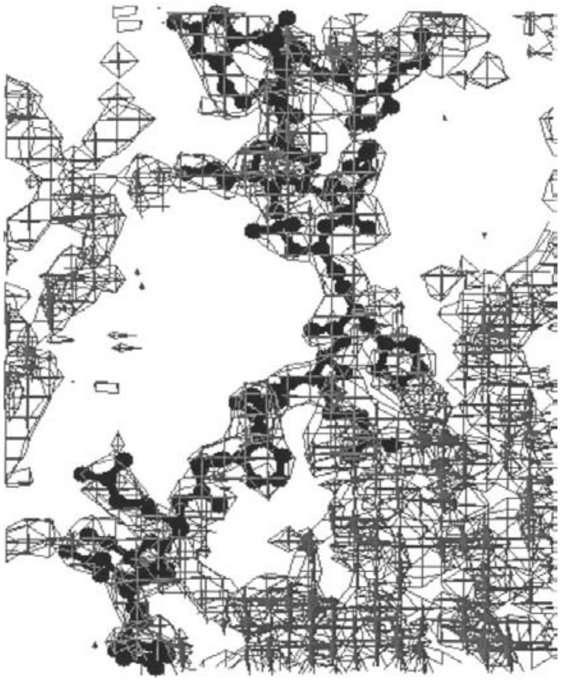

Figure 3
Same as Fig. 2, except that the electron density shown is that reconstructed by Sim's weighted difference Fourier prescription. The isosurface represented by the wire mesh is that corresponding to a density of 1.3 times the standard deviation of the electron-density map above the mean density. Also shown is a ball-and-stick figure representing the residues 1–18.

The starting point of the theory is the fact that, in Boltzmann's expression for the entropy, S, of a distribution $\{u_j\}$, namely

$$S[\{u_j\}] = k \ln \Omega[\{u_j\}], \tag{15}$$

where k is Boltzmann's constant, the number of microstates per macrostate, Ω, is proportional to the probability (P) of the distribution. Consequently,

$$P[\{u_j\}] \propto \exp S[\{u_j\}]. \tag{16}$$

Thus the most probable distribution $\{u_j\}$ corresponds to that which maximizes S. A convenient form for the entropy, which is equivalent to Boltzmann's expression above, is Gibbs's form (Landau & Lifshitz, 1980):

$$S[\{u_j\}] = -\sum_j u_j \ln [u_j/(eq_j)], \tag{17}$$

where e is the base of the natural logarithms and $\{q_j\}$ the best prior guess of the optimum distribution $\{u_j\}$ (which we could term the *measure* of the distribution). By differentiating S with respect to u_i (where i is a particular one of the set of indices $\{j\}$), it is easy to show that the distribution $\{u_j\}$ that maximizes S is the trivial one that is identical to $\{q_j\}$.

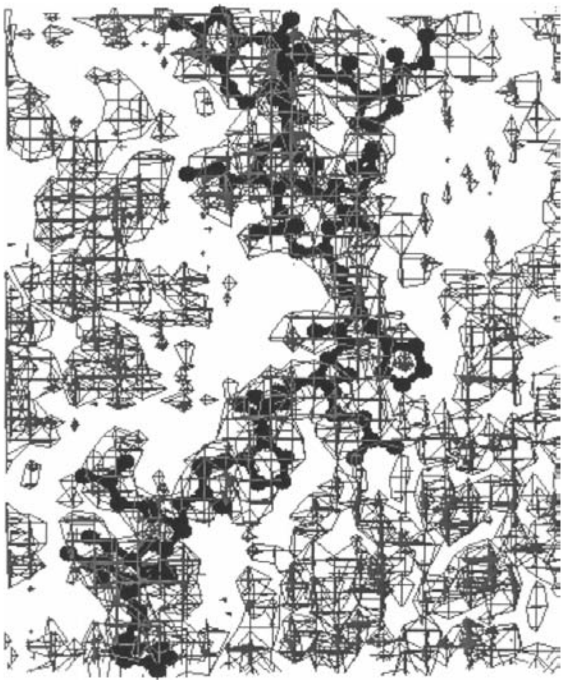

Figure 4
Same as Fig. 2, except that the electron density shown is that recovered by density-modified difference Fourier synthesis. The isosurface represented by the wire mesh is that corresponding to an electron density of 1.1 times the standard deviation of the density above the mean.

For our problem of finding the most probable electron distribution $\{u_j\}$ consistent with the experimental data, we need to constrain the distribution by the method of Lagrange multipliers. In the case of the structure completion problem, we identify $\{u_j\}$ with our best guess of the distribution $\{u_j^{(n)}\}$ of the unknown part of a unit cell at step n of an iterative algorithm. We identify the measure $\{q_j\}$ with our estimate $\{u_j^{(n-1)}\}$ of the electron distribution at the previous iteration. We seek to maximize the functional

$$Q[\{u_j^{(n)}\}] = -\sum_j u_j^{(n)} \ln\left[\frac{u_j^{(n)}}{e u_j^{(n-1)}}\right] - \frac{\lambda'}{2}\sum_{\mathbf{g}} \frac{|O_{\mathbf{g}}^{(n)} - T_{\mathbf{g}}^{(n-1)}|^2}{\sigma_{\mathbf{g}}^2}, \tag{18}$$

where the first term on the RHS is Gibbs's expression for the entropy of the distribution $\{u_j^{(n)}\}$ with respect to that, $\{u_j^{(n-1)}\}$, from the previous iteration. The second term on the RHS constrains the calculated structure factors

$$O_{\mathbf{g}}^{(n)} = \sum_j u_j^{(n)} \exp(i\mathbf{g}\cdot\mathbf{r}_j) \tag{19}$$

from the unknown part of the structure to be consistent with the experimental data, represented by a set of *target* structure factors

$$T_{\mathbf{g}}^{(n-1)} = \{|F_{\mathbf{g}}| \exp(i\varphi_{\mathbf{g}}^{(n-1)}) - R_{\mathbf{g}}\}, \tag{20}$$

where $\sigma_{\mathbf{g}}$ is the estimated uncertainty in the measured structure-factor amplitude $|F_{\mathbf{g}}|$,

$$\varphi_{\mathbf{g}}^{(n-1)} = \arg(R_{\mathbf{g}} + O_{\mathbf{g}}^{(n-1)}) \quad \text{if } n \geq 2, \tag{21}$$

and λ' is a Lagrange multiplier.

Several different types of constraints have been proposed for maximum-entropy applications to X-ray crystallography. An excellent review of different forms are found in a paper by Wilkins (1983*a*). He makes a distinction between *hard constraints* in which the electron distribution is constrained by the amplitudes and/or known phases of *individual* reflections, and *weak constraints* in which the constraints are a single function with contributions from all the Bragg reflections. In (18), we use the form of weak constraint proposed by Collins (1982). This has the advantages that (*a*) experimental errors in individual structure factors are less likely to lead to spurious detail in reconstructed electron distributions, and (*b*) the number of Lagrange multipliers to be determined is much smaller (in this case only one, if normalization is performed separately). In contrast, the hard constraints employed by Bricogne (1984) requires the determination of $N+1$ Lagrange multipliers where N is the number of constraining reflections.

Q may be maximized by requiring that

$$\frac{\partial Q}{\partial u_i^{(n)}} = 0 \quad \forall\, i. \tag{22}$$

The differentiation of the entropy term in (18) is straightforward enough; that of the constraint term may be performed by writing $|O_{\mathbf{g}}^{(n)} - T_{\mathbf{g}}^{(n-1)}|^2$ as $\{O_{\mathbf{g}}^{(n)} - T_{\mathbf{g}}^{(n-1)}\}$ times its complex conjugate and noting that $O_{\mathbf{g}}^{(n)}$ depends on $u_i^{(n)}$, but not $T_{\mathbf{g}}^{(n-1)}$. After some algebra and making use of the inverse transform of (19), namely

$$u_i = (1/N)\sum_{\mathbf{g}} O_{\mathbf{g}}^{(n)} \exp(-i\mathbf{g}\cdot\mathbf{r}_i) = (1/N)\sum_{\mathbf{g}} O_{\mathbf{g}}^{(n)*} \exp(i\mathbf{g}\cdot\mathbf{r}_i) \tag{23}$$

(the last equality follows from Friedel's law, $O_{\mathbf{g}} = O^*_{-\mathbf{g}}$, and replacing the dummy index $-\mathbf{g}$ by $\mathbf{g}$ under the summation over $\mathbf{g}$); this leads to the 'single voxel' equations

$$\ln[u_i^{(n)}/u_i^{(n-1)}] = -\lambda\{u_i^{(n)} - t_i^{(n-1)}\}, \tag{24}$$

where $\lambda = \lambda' N/\langle\sigma_{\mathbf{g}}^2\rangle$ if we replace the individual variances $\sigma_{\mathbf{g}}^2$ by their mean value, and

$$t_i^{(n-1)} = (1/N)\sum_{\mathbf{g}} T_{\mathbf{g}}^{(n-1)} \exp(-i\mathbf{g}\cdot\mathbf{r}_i) \tag{25}$$

is a *target function* consisting of the inverse Fourier transform of $T_{\mathbf{g}}^{(n-1)}$. For more general constraints, Wilkins (1983*a*,*b*) had earlier derived analogous 'single pixel' equations by ignoring the off-diagonal terms of a Hessian matrix in a Taylor-series expansion of the constraint functions. We stress that, for the particular form of constraints in (18), the only approximation in the derivation of the single voxel equations is the replacement of the individual variances $\sigma_{\mathbf{g}}^2$ by their mean. Hence,

$$u_i^{(n)} = u_i^{(n-1)} \exp[-\lambda\{u_i^{(n)} - t_i^{(n-1)}\}]. \tag{26}$$

This is an *implicit* relation for $u_i^{(n)}$ in terms of $u_i^{(n-1)}$ and $t_i^{(n-1)}$. It can be written as an *explicit* equation for $u_i^{(n)}$ by substituting $u_i^{(n-1)}$ for $u_i^{(n)}$ on the RHS. This substitution would be justified only if λ were chosen small enough that

$$|\delta u_i^{(n)}| \ll |u_i^{(n)} - t_i^{(n-1)}|, \tag{27}$$

where

$$\delta u_i^{(n)} = u_i^{(n)} - u_i^{(n-1)}. \tag{28}$$

Note that if λ were small enough it would be possible also to truncate the series expansion of the exponential on the RHS of (26) to approximate this equation by

$$\delta u_i^{(n)} = -\lambda u_i^{(n-1)}\{u_i^{(n)} - t_i^{(n-1)}\} \tag{29}$$

from which it follows that condition (27) is equivalent to the requirement that $|\lambda u_i^{(n-1)}| \ll 1$ or, alternatively, $\lambda \ll 1/u_i^{(n-1)}$ $\forall\, i$. This can be ensured by choosing

$$\lambda \ll 1/u_{\max}^{(n-1)}, \tag{30}$$

where $u_{\max}^{(n-1)}$ is the maximum value of the distribution $\{u_i^{(n-1)}\}$. It should be noted that a λ chosen according to this prescription would almost certainly also justify the truncation of the series expansion of the exponential in (26) that leads to (29) (since $u_i^{(n-1)}$, $u_i^{(n)}$ and $t_i^{(n-1)}$ are all similar in magnitude by construction), so the argument is self-consistent. Thus we may replace (26) by the following *explicit* recursion relation:

$$u_i^{(n)} = u_i^{(n-1)} \exp[-\lambda\{u_i^{(n-1)} - t_i^{(n-1)}\}] \tag{31}$$

so long as λ satisfies (30). The algorithm is initiated by defining a starting distribution $\{u_i^{(0)}\}$ for the sought electron distribution, and also $\{t_i^{(0)}\}$ for the 'target' function. These distributions need to be different, otherwise the argument of the exponential in (31) will be zero and the sought distribution $\{u_i\}$ will not be updated. Our procedure for constructing these initial distributions from a standard crystallographic computer program applied to the experimental data is described in the next section. The initial electron distribution $\{u_i^{(0)}\}$ is designed to resemble a fuzzy molecular envelope using experimental data alone. That distribution is updated at each iteration only from (31) and from a re-normalization to the expected total number of electrons.

The construction of $\{u_i^{(0)}\}$ produces a positive-definite distribution. The exponential in (31) ensures that the recursion relation can never produce negative values at any voxel at any subsequent iteration. This process of *exponential modeling* (Collins & Mahar, 1983; Carter & Xiang, 1997) automatically satisfies the physical constraint of positivity of the electron distribution.

It is our experience that, provided the parameter λ is chosen consistent with (27), the algorithm invariably improves the initial phase estimates $\varphi_{\mathbf{g}}^{(0)}$ during the course of the first several iterations (as monitored by a correlation coefficient between the distribution $\{u_i^{(n)}\}$ and the corresponding electron distribution of the solved structure). This is consistent with the results of Collins (1982) in his work on the phase refinement of an entire protein. During this phase of the iterations, the electron distribution $\{u_i\}$ approaches that of the target function $\{t_i\}$, which in turn is constrained by the reciprocal-space values of its Fourier coefficients $T_{\mathbf{g}}$ (20). At the same time, the algorithm does not allow the distribution to rapidly stray too far away from the shape of its initial molecular envelope $\{u_i^{(0)}\}$. It is this competition between reciprocal-space constraints and bias towards the shape of the molecular envelope that gives the algorithm its remarkable phasing and map improvement capabilities. It should be noted that there is no strict molecular envelope constraint and the algorithm is able to correct errors in that estimate if so dictated by the reciprocal-space constraints.

Nevertheless, if the iterations are allowed to proceed too far, the electron map may lose memory of the initial molecular envelope. Therefore, it is important to halt the iterations at an optimal point. Fortunately, as we describe in the next section, even in the case of an unknown missing molecular fragment, it is possible to determine this point from the minimum of the cross-validation measure termed by Brünger (1992) a *free R factor*.

We shall next describe an application of this algorithm to the structure-completion problem from the experimental data of α-dendrotoxin.

6. Structure completion of α-dendrotoxin by exponential modeling

In order to initialize the algorithm (31), starting distributions $\{u_i^{(0)}\}$ and $\{t_i^{(0)}\}$ must be set up. In our earlier work with synthetic data (Saldin *et al.*, 1997), we equated the target function $t_i^{(0)}$ to the RHS of (5), namely the unweighted difference Fourier estimate. Better starting distributions $\{t_i^{(0)}\}$ are either the SWDF map (8) or the DMDF one (13). When attempts were made to recover the deleted residues 1–18 of α-dendrotoxin, we found that, if we started with a SWDF map (with correlation coefficient with the 'exact' map of $C = 0.41$), the algorithm improved this map considerably, as characterized by a value of $C = 0.61$. For the same problem, starting instead with the DMDF map (with $C = 0.43$) gave rise to an even better final map (with $C = 0.70$). Therefore, in the following, we will describe in detail only work in which $\{t_i^{(0)}\}$ is equated to the better DMDF starting distribution (13).

Generally, in protein crystallography the structure factor of forward-scattered X-rays, corresponding to the $\mathbf{g} = \mathbf{0}$ reciprocal-lattice vector is not measured. However, its value is simply the total number of electrons in the unit cell. These consist both of protein electrons as well as those of its water solvent, whose density is usually estimated at ~0.32 electrons Å^{-3}. Thus, the value of $F_{\mathbf{0}}$ is just the sum of the number $N_{\rm protein}$ of protein electrons and $N_{\rm solvent}$ of those of the solvent. In a difference Fourier formula, the corresponding structure factor $R_{\mathbf{0}}$ of the partial structure is equal to the number $N_{\rm partial}$ of electrons in that partial structure. Thus, strictly, a difference Fourier formula like (8) or (13) should have a $\mathbf{g} = \mathbf{0}$ Fourier coefficient equal to $N_{\rm protein} + N_{\rm solvent} - N_{\rm partial}$ and should recover not only the electron distribution of the missing residues but also that of the solvent. Since the focus of structure completion work is the recovery of the missing residues rather than that of the solvent, it makes sense to simply drop the term $N_{\rm solvent}$ and take the $\mathbf{g} = \mathbf{0}$ Fourier coefficient of a difference Fourier formula equal to $N_{\rm protein} - N_{\rm partial}$. The quantity $N_{\rm protein}$ is generally known from prior biochemical analysis; $N_{\rm partial}$ is known by hypothesis. The effect of this choice of the $\mathbf{g} = \mathbf{0}$ Fourier coefficient is to subtract an electron density equal to the solvent density times the ratio (solvent volume)/(total volume) of the unit cell from all the recovered electron density. This increases the (protein density)/(solvent density) ratio in the difference Fourier map, a desirable feature when attempting to recover the molecular envelope of the missing residues. Since, in our exponential modeling algorithm, the target function $t_i^{(n)}$ is also constructed from a difference Fourier formula, the same choice of the $\mathbf{g} = \mathbf{0}$ Fourier coefficient is appropriate for its construction also.

In order to monitor the improvement in the phasing of the reflections, we set aside a randomly chosen subset (we chose 7.5%) of the structure factors, which were not used in the phasing algorithm. Following Brünger (1992, 1993, 1997), this subset is known as the *test set*, T. The remaining reflections form the working set W. Thus we took

$$t_i^{(0)} = (1/N) \sum_{\mathbf{g}\in W} [w_{\mathbf{g}}^{(\rm DM)} |F_{\mathbf{g}}| \exp\{i\varphi_{\mathbf{g}}^{(\rm DM)}\} - R_{\mathbf{g}}] \exp(-i\mathbf{g}\cdot\mathbf{r}_i). \tag{32}$$

As for the initial distribution, $\{u_i^{(0)}\}$, we take this to be a low-pass filtered version of $\{t_i^{(0)}\}$, by defining

$$u_i^{(0)} = (1/N) \sum_{\mathbf{g}\in W} [w_{\mathbf{g}}^{(\mathrm{DM})}|F_{\mathbf{g}}| \exp(i\varphi_{\mathbf{g}}^{(\mathrm{DM})}) - R_{\mathbf{g}}] \exp(-i\mathbf{g}\cdot\mathbf{r}_i) \times \exp[-\tfrac{1}{2}(g\delta)^2], \qquad (33)$$

where δ is the real-space resolution and g is the magnitude of the reciprocal-lattice vector $\mathbf{g}$. This imposes a Gaussian envelope in reciprocal space. As suggested by Collins (1982), negative and small positive electron densities were eliminated by replacing all values of $u_i^{(0)} < u_{\max}^{(0)}/100$ by $u_{\max}^{(0)}/100$, where $u_{\max}^{(0)}$ is the maximum value of the distribution $\{u_i^{(0)}\}$.

We began with a value $\delta = 10$ Å. Subsequent evaluations of the target function $t_i^{(n-1)}$ from (25) were also from Fourier transforms of target amplitudes $T_{\mathbf{g}}^{(n-1)}$ with $\mathbf{g} \in W$. Thus the entire phasing algorithm used information from just 92.5% of randomly chosen structure factors from the entire measured set. Owing to the exponentiation in (31), structure factors $O_{\mathbf{g}}^{(n)}$ evaluated from (19) will include reflections $\mathbf{g} \in T\ \forall n > 0$. We define a *free R factor* at iteration n by

$$R_{\mathrm{free}}^{(n)} = \frac{\sum_{\mathbf{g}\in T} ||F_{\mathbf{g}}| - |R_{\mathbf{g}} + O_{\mathbf{g}}^{(n)}||}{\sum_{\mathbf{g}\in T} |F_{\mathbf{g}}|}, \qquad (34)$$

which measures the agreement between the calculated structure-factor amplitudes $|R_{\mathbf{g}} + O_{\mathbf{g}}^{(n)}|$ belonging to the test set with the corresponding measured quantities $|F_{\mathbf{g}}|$ belonging to the same set. As Brünger (1992, 1993, 1997) has pointed out, this free R factor has the remarkable property of being able to monitor the quality of the *phasing* of the working set of reflections W, even in the case of an unknown structure.

Our experience bears this out: for the recovery of deleted residues 1–18 (30% of the structure), Table 1 shows the variation of $R_{\mathrm{free}}^{(n)}$ with iteration number n. $R_{\mathrm{free}}^{(n)}$ is seen to reach a minimum when $n = 10$ and to rise after that. The validity of $R_{\mathrm{free}}^{(n)}$ as a monitor of the quality of the electron distribution may be judged by a comparison with the simultaneous variation with n of the correlation coefficient C_W between the electron distribution $\{u_i^{(n)}\}$ calculated from reflections of just the working set W and that of residues 1–18 of the published structure (Skarzynski, 1992). It will be noted that C_W increases until the same iteration number 10, after which it decreases. This indicated that, not only is the best electron distribution obtained after 10 iterations, but also that this best map may be identified by the minimum in the free R factor, whose calculation does not require any prior knowledge of the true structure. Let us define the iteration number that gave the best map in this cycle as m. In this case therefore, $m = 10$. The corresponding column in Table 1 is highlighted in bold-face type.

We also found that this map may be improved further by a new cycle of iterations of the phasing algorithm. The initial map of this new cycle is defined by

$$u_i^{(0)} = (1/N) \sum_{\mathbf{g}\in W} [|F_{\mathbf{g}}| \exp(i\varphi_{\mathbf{g}}^{(m)}) - R_{\mathbf{g}}] \exp(-i\mathbf{g}\cdot\mathbf{r}_i) \times \exp[-\tfrac{1}{2}(g\delta)^2], \qquad (35)$$

where the phase of $F_{\mathbf{g}}$ is taken as that of the best map from the previous iteration, *i.e.* $\varphi_{\mathbf{g}}^{(m)}$, and the map effectively convolved with a Gaussian of reduced width of $\delta = 5$ Å. When the iterations were resumed, the quality of the map improved further, again monitored by $R_{\mathrm{free}}^{(n)}$ and C. The former quantity is approximately constant for 3 iterations and rises thereafter. For both sets of iteration cycles, the optimal stopping point is seen to be the last iteration before a rise in $R_{\mathrm{free}}^{(n)}$. That optimal column is also distinguished by bold characters.

Thus the algorithm was not only able to increase the correlation with the 'exact' electron distribution of the solved structure but also the cross-validation technique of Brünger (1992, 1993, 1997) gives an accurate independent indication of the best map and provides a criterion for halting the iterations when this best map is obtained. A slight disadvantage of this procedure is the fact that the map is calculated with just the W set of reflections that constitute less than all the available diffraction data. Therefore, we used the algorithm to evaluate also an electron distribution using the full set of the available reflections for the same number of iterations as the optimum for the W set. We found that the maximum correlation C_F of the full set of Bragg reflections between the new reconstructed electron distribution and that of the model distribution was also reached at the same iteration numbers for each of the iteration cycles, as illustrated in the final rows of Table 1. Thus, it appears that Brünger's free R factor can be used to determine the optimum number of iterations per resolution cycle and then a calculation of the best map can be re-performed using all the available Bragg reflections, terminating each cycle at the optimum number of iterations determined by the free R factor. By this means, a further improvement of the final electron distribution was found, as quantified by a value of $C_F = 0.70$, as may be seen from Table 1.

The above procedure was repeated to recover residues 1–30 (50% of the structure), after they had been deleted from the partial structure of Fourier coefficients $R_{\mathbf{g}}$. As indicated in Table 2, the minimum in $R_{\mathrm{free}}^{(n)}$ occurred at iteration number $n = 12$, which also corresponded to a maximum of the corresponding correlation coefficient C_W during the first cycles with $\delta = 10$ Å. Again using (35) to re-start a new set of iterations with $\delta = 5$ Å, R_{free} remained approximately constant for a further 3 iterations before rising. Once again it is seen that the last iteration before the rise in R_{free} corresponds to the maximum of both C_W and C_F. The value of C_F, characterizing the quality of the best reconstructed map of residues 1–30, is also seen to be 0.70.

In our test calculation with the recovery of residues 1–18 of α-dendrotoxin (30% of the structure), exponential modeling was able to recover a map of those residues (Fig. 5) of much higher quality (of correlation with the 'exact' structure $C = 0.70$) than those from the SWDF synthesis ($C = 0.41$) or from the DMDF phases and weights in (13) (for which $C = 0.43$). Comparison of Fig. 5 with the 'exact' map of Fig. 2 shows striking agreement and a clear identification of the continuous tube of electron density associated with the polypeptide chain.

Even when the recovery was attempted of half of the structure (residues 1–30), the corresponding correlation coefficient was found to be of the same value ($C = 0.70$),

Table 1
Recovery of residues 1–18.

Resolution of starting map, $\{u_i^{(0)}\}$: $\delta = 10$ Å

n	1	2	3	4	5	6	7	8	9	**10**	11	12	13
$R_{\rm free}^{(n)}$	32.08	17.99	10.61	6.73	4.40	2.73	1.54	0.89	0.61	**0.51**	0.52	0.54	0.57
C_W	0.21	0.28	0.35	0.42	0.47	0.52	0.57	0.60	0.61	**0.61**	0.60	0.59	0.58
C_F	0.20	0.27	0.35	0.42	0.48	0.53	0.57	0.60	0.63	**0.64**	0.64	0.63	0.63

Resolution of re-start map, $\{u_i^{(0)}\}$: $\delta = 5$ Å

n	1	2	**3**	4	5	6
$R_{\rm free}^{(n)}$	0.41	0.40	**0.40**	0.41	0.43	0.44
C_W	0.66	0.67	**0.67**	0.67	0.66	0.65
C_F	0.68	0.69	**0.70**	0.70	0.70	0.70

again considerably in excess of that of the SWDF map ($C = 0.42$) or the DMDF map ($C = 0.50$)

In the case of each of the electron distributions illustrated in our paper, the isosurface contour level was chosen to be the highest that displays a more-or-less continuous tube of electron density around the recovered residues. On comparing Figs. 2, 3, 4 and 5, it is seen that, as expected, the better the reconstruction of the amino acid chain, the higher is this contour level in relation to the mean density.

Further insight into the quality of the reconstructed electron maps of residues 1–18 may be obtained by examining Fig. 6, which depicts the variation of the correlation coefficient for a number of resolution shells characterized by their mean values of $\{\sin(\theta)/\lambda''\}^2$ (where θ is half the angle of scattering and λ'' the wavelength of the X-rays) for the Sim-weighted and density-modified difference Fourier maps, and the map recovered by the exponential modeling algorithm. The last-named map has a significantly higher correlation with the 'exact' map than either of the other two. Also very interesting is the fact that, although the SWDF map is superior to the corresponding density-modified one over most of the resolution range, consistent with the appearances of the maps of Figs. 3 and 4, the DMDF map is better in the low-resolution region up to a value of the abscissa of about 0.01. This superiority of the map that uses the density-modified phases and weights in the low-resolution regime is not surprising since the density-modification process (Cowtan & Main, 1998) involves the determination of a low-resolution molecular envelope. This is also the reason for our choice of initial low-pass filtered electron distribution $\{u_i^{(0)}\}$ to be of the form (33) rather than from the SWDF map (8).

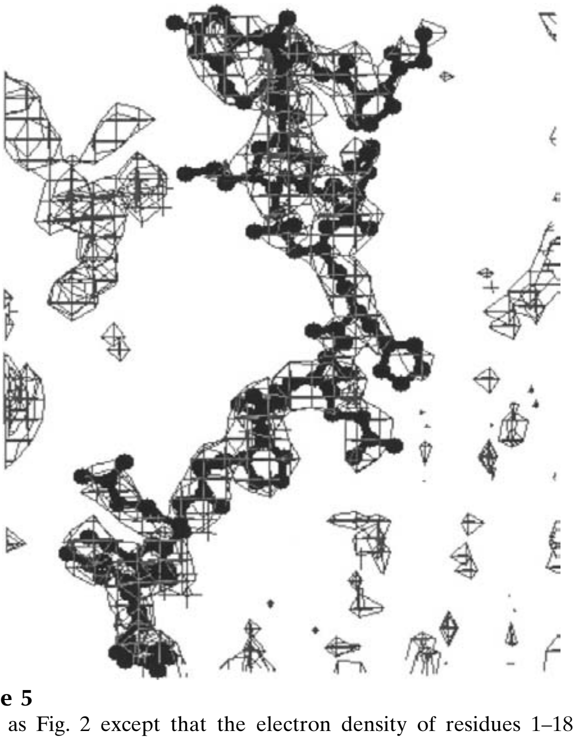

Figure 5
Same as Fig. 2 except that the electron density of residues 1–18 was reconstructed by the exponential modeling algorithm described in the text. The wire-mesh isosurface corresponds to an electron density of 2.0 times the standard deviation of the density above the mean. This electron-density distribution is seen to give a much clearer indication of the three-dimensional configuration of the deleted residues 1–18 than that of Figs. 2 or 3. Much of the electron density not enclosing the ball-and-stick figure represents parts of symmetry-related molecules in the crystal. Note the striking similarity with the 'exact' map of Fig. 2.

7. Discussion

Our algorithm for structure completion is similar to one proposed by Collins (1982) for improving the resolution of an initial map of the electron density of an entire protein based on experimental phases, such as those obtained by the MIR technique. Apart from addressing the different problem of structure completion, the new feature of our scheme is the provision for *cross validation*, which determines the optimal stopping point for the recursion relation.

The framework for the derivation of the algorithm is the maximum-entropy formalism of Jaynes (1957). The centerpiece of this method as applied to protein crystallography is the definition of a functional, Q (18), of the electron distribution that consists of a weighted sum of two terms: (*a*) the entropy S of the electron distribution $\{u_i\}$ *relative* to a measure $\{q_i\}$ that represents an *a prori* estimate of $\{u_i\}$ and (*b*) another functional that constrains the electron density to the experimentally observed amplitudes of Bragg reflections and to initial phase estimates. Collins employs the *weak constraint* of a χ^2 statistic that results in the relative weights of these two functional contributions to Q being controlled by a *single* Lagrange multiplier λ'. The particular form of χ^2 chosen by Collins, which makes this functional the logarithm of a *likelihood function* (Sivia, 1996), also makes it a *quadratic* functional of $\{u_i\}$. Functional differentiation of Q with respect to the electron distribution yields an equation (26) to be satisfied by that distribution. The quadratic nature of χ^2 with respect to $\{u_i\}$ results in an *implicit* equation for $\{u_i\}$. Solving that

Table 2
Recovery of residues 1–30.

Resolution of starting map, $\{u_i^{(0)}\}$: $\delta = 10$ Å													
n	1	2	3	4	5	6	7	8	9	10	11	**12**	13
$R_{\rm free}^{(n)}$	105.99	54.03	29.48	17.78	11.53	7.23	4.19	2.33	1.30	0.76	0.54	**0.50**	0.53
C_W	0.15	0.20	0.27	0.34	0.39	0.44	0.49	0.54	0.57	0.60	0.61	**0.61**	0.60
C_F	0.15	0.20	0.27	0.34	0.40	0.45	0.50	0.55	0.59	0.62	0.64	**0.65**	0.65

Resolution of re-start map $\{u_i^{(0)}\}$: $\delta = 5$ Å						
n	1	2	**3**	4	5	6
$R_{\rm free}^{(n)}$	0.49	0.49	**0.49**	0.50	0.52	0.53
C_W	0.65	0.66	**0.66**	0.66	0.66	0.66
C_F	0.68	0.69	**0.70**	0.70	0.70	0.70

equation for $\{u_i\}$ for a *fixed* $\{q_i\}$ while evaluating λ' by some auxiliary constraint equation would yield the conventional maximum-entropy distribution for $\{u_i\}$.

This is essentially the method followed by Bricogne (1984), except that he employed *hard* constraints for the individual Bragg reflections (which give rise to $N+1$ Lagrange multipliers, where N is the number of measured reflections) which were *linear* in the electron distribution. A similar functional differentiation then yields *explicit* equations for the maximum-entropy electron distribution, with the Lagrange multipliers determined by $N+1$ auxiliary constraint equations.

At this point, the Collins algorithm (and ours) departs from the conventional maximum-entropy treatment. Instead of being treated as a constant, the *measure* $\{q_i\}$ in the relative entropy S is continually updated during the iterations to find $\{u_i\}$. For this reason, we refrain from using the term *maximum entropy* to describe the algorithm but rather use instead the term subsequently coined by Collins & Mahar (1983), *exponential modeling*. For the problem we address in this paper, we found this procedure actually more effective at leading to an improved electron distribution than a conventional maximum-entropy solution in which $\{q_i\}$ is kept fixed at its initial estimate.

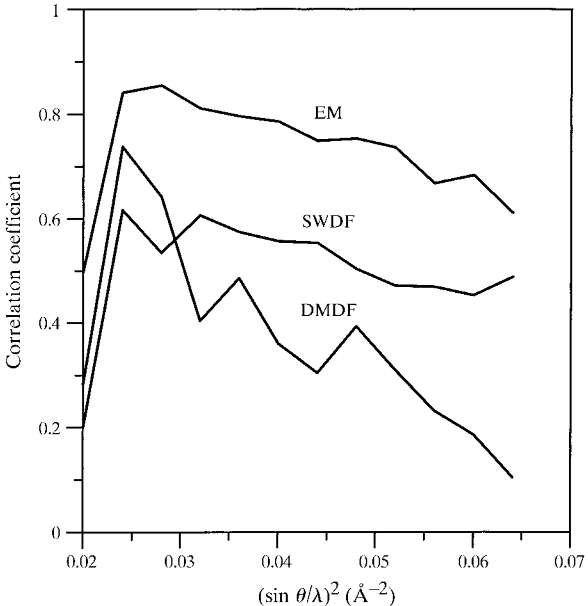

Figure 6
Linear correlation coefficients comparing the 'exact' electron density of residues 1–18 of α-dendrotoxin with that of the same residues recovered by our exponential modeling algorithm (EM), the Sim-weighted difference Fourier method (SWDF) and the density-modified difference Fourier method (DMDF) as a function of resolution shell. The quality of the exponential modeling reconstruction is clearly superior to that of either of the other two methods over the entire resolution range. The DMDF maps are better than the SWDF ones at low resolution but less good at high resolution.

One reason for the effectiveness of the method becomes apparent on examination of the recursion relation (31). In this expression, $\{t_i^{(n-1)}\}$ is the Fourier transform of the reciprocal-space estimate of the unknown electron distribution with the best estimate of the phases of the measured structure-factor amplitudes from the previous iteration. If the estimate $u_i^{(n-1)}$ of the sought distribution at the previous iteration is less than the corresponding value $t_i^{(n-1)}$ of this *target function*, the estimate of u_i at iteration n will be greater than that at the previous iteration and *vice versa*. Thus the sought distribution tends towards a compromise between the initial real-space estimate represented by $\{u_i^{(0)}\}$ and consistency with reciprocal-space constraints contained in $\{t_i\}$. This is what gives the algorithm its phasing or equivalently its electron map improvement capabilities.

It should also be noted that, in our scheme, the quantity λ plays the role of a relaxation parameter that ensures the proper operation of the recursion relation (31). If every iteration yields a distribution $\{u_i^{(n)}\}$ that more accurately represents the electron distribution sought, this will also yield better phase estimates $\{\varphi_{\bf g}^{(n)}\}$ of the measured Bragg reflections. The complementary improvements in real and reciprocal space act as a kind of feedback loop. This is a feature shared with the density-modification method (Cowtan & Main, 1998), the *MESF* algorithm of Xiang *et al.* (1993) and even the Gerchberg–Saxton algorithm (Gerchberg & Saxton, 1972) of image processing. A representative amplitude–phase diagram indicating the relationships amongst the different structure factors is illustrated in Fig. 7.

Our method also bears some formal analogy to Bricogne's (1993) Bayesian methods for *phase extension*, where guessed values of the phases of an initial subset H of reflections is used in a (different) exponential modeling algorithm to estimate the amplitudes of a neighboring subset K of Bragg reflections. In Bricogne's method, the quality of the initial phase guesses is monitored by a quantity termed a *log-likelihood gain*, which compares the degree of agreement between the predicted and measured amplitudes of the K subset. In our scheme, we start with approximate phase estimates of *all* measured reflections but then randomly choose just 92.5% of them as input to our exponential modeling algorithm. The remaining 7.5% plays essentially the same *cross-validation* role as Bricogne's K subset. In our case, the step analogous to Bricogne's *phase hypothesis generation* is performed automatically by our algorithm, and our cross-validating *free R factor* merely determines the iteration number of the most appropriate stopping point. Since we have found from experience that, if all the experimental data are then included in the algorithm, an even better map is found after the same number of iterations, we recommend using all the data at this final step (thus using the cross-validation procedure only to determine the optimal iteration number for the full data set).

8. Conclusions

We have proposed an exponential modeling algorithm to address the problem of structure completion in protein X-ray crystallography, particularly where the unknown part of the structure may be a significant proportion of the whole protein, as may be encountered in the method of molecular replacement, for example.

We have demonstrated the effectiveness of our algorithm with a test where we attempted to recover the electron distribution of a significant portion of a 59-residue protein, α-dendrotoxin, from experimental structure factors and a knowledge of as little as half of the structure. Neither a Sim-weighted nor a density-modified difference Fourier map was able to clearly identify the chain of the missing residues. In contrast, our algorithm was able to improve these maps to the point where a continuous tube of electron density representing the shape of the corresponding polypeptide chain stood out clearly from surrounding noise. Strong confirmation of these observations was found in the more objective measure of a coefficient of correlation with the known electron distribution of the missing residues.

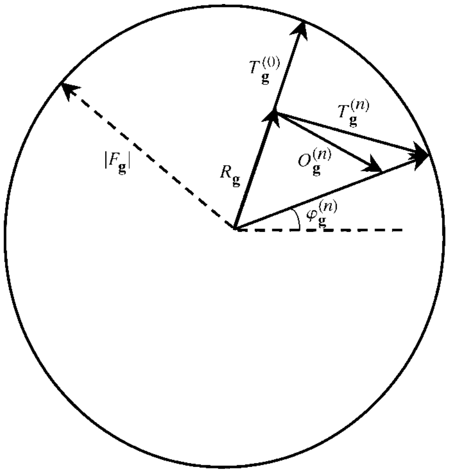

Figure 7
Amplitude–phase diagram indicating the relationships amongst the various component structure factors of reciprocal-lattice vector $\mathbf{g}$. The circle has a radius of $|F_{\mathbf{g}}|$, the measured amplitude of Bragg reflection $\mathbf{g}$. $R_{\mathbf{g}}$ represents the structure factor of the known molecular fragment. This is known in both amplitude (length) and phase (angular separation from the dashed line). The (unweighted) difference Fourier estimate of the structure factor of the missing residues is represented by the vector $T_{\mathbf{g}}^{(0)}$, which has the same phase (direction) as $R_{\mathbf{g}}$. $O_{\mathbf{g}}^{(n)}$ is the estimate of the same structure factor at the nth iteration of the exponential modeling algorithm. Since the end of the vector sum of $R_{\mathbf{g}}$ and $O_{\mathbf{g}}^{(n)}$ will not in general lie on the circumference of the circle, the length of this vector is adjusted to the circle radius. The target structure factor $T_{\mathbf{g}}^{(n)}$ of the missing residues is then constructed such that when added vectorially to $R_{\mathbf{g}}$ it is equal in both amplitude $|F_{\mathbf{g}}|$ and phase ($\varphi_{\mathbf{g}}^{(n)}$) to the new estimate $F_{\mathbf{g}}^{(n)}$ of the structure factor of the entire protein. The Fourier transform of the target structure factors $\{T_{\mathbf{g}}^{(n)}\}$ forms the next estimate of the target function $\{t_i\}$ used to evaluate the next estimate of the distribution $\{u_i\}$ of the missing electrons from the recursion relation (31) of the text. In turn the Fourier transform of this distribution forms the estimate of $O_{\mathbf{g}}$ for the next iteration $(n+1)$.

The algorithm may also be thought of as one that progressively improves the phase estimates of the measured Bragg reflections by alternate cycles between real and reciprocal space. Each new real-space distribution gives rise to a new set of phases on Fourier transformation. At each reciprocal-space step, the amplitudes associated with those new phases are constrained by the measured (and scaled) experimental values and an estimate of the total number of electrons in the missing part of the protein. The method naturally incorporates the necessary positivity constraint on the electron density. It improves the phases of experimental Bragg reflections by the steps of *phase hypothesis generation* (performed automatically according to a prescription) followed by *cross validation* by means of a free R factor.

DKS is grateful for financial support from the US National Science Foundation (grant no. DMR-9815092). He acknowledges stimulating discussions with Wolfgang Moritz and Abraham Szöke.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.
Bricogne, G. (1991). *Maximum Entropy in Action*, edited by B. Buck & V. A. Macaulay, pp. 187–216. Oxford University Press.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24–36.
Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.
Carter, C. W. & Xiang, S. (1997). *Methods Enzymol.* **277**, 79–109.
Cochran, W. (1951). *Acta Cryst.* **4**, 408–401.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Collier, R. J., Burckhardt, C. B. & Lin, L. H. (1971). *Optical Holography.* San Diego: Academic Press.
Collins, D. M. (1982). *Nature (London)*, **298**, 49–51.
Collins, D. M. & Mahar, M. C. (1983). *Acta Cryst.* A**39**, 252–256.
Cowtan, K. & Main, P. (1998). *Acta Cryst.* D**54**, 487–493.
Drenth, J. (1994). *Principles of Protein X-ray Crystallography.* New York: Springer-Verlag.
Fitzgerald, P. M. (1994). In *From First Map to Final Model.* Proceedings of the 1994 CCP4 Study Weekend. CLRC Daresbury Laboratory, England.
Gabor, D. (1948). *Nature (London)*, **161**, 777–778.
Gerchberg, R. W. & Saxton, W. O. (1972). *Optik (Stuttgart)*, **35**, 237–246.
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 287–307.

Gull, S. F. & Daniell, G. J. (1978). *Nature (London)*, **272**, 686–690.

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992) *Acta Cryst.* A**48**, 851–858.

Jaynes, E. T. (1957). *Phys. Rev.* **106**, 620–630.

Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.

Landau, L. D. & Lifshitz, E. M. (1980). *Statistical Physics.* New York: Pergamon Press.

Leahy, D. J., Hendrickson, W. A., Aukhil, I. & Erickson, H. P. (1992). *Science*, **258**, 987–991.

Maalouf, G. J., Hoch, J. C., Stern, A. S., Szöke, H. & Szöke, A. (1993). *Acta Cryst.* A**49**, 866–871.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in FORTRAN*, 2nd ed., p. 630. Cambridge University Press.

Read, R. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. (1997). *Methods Enzymol.* **277**, 110–128.

Rossmann, M. G. & Blow, D. M. (1962) *Acta Cryst.* **15**, 24–31.

Saldin, D. K., Chen, X., Kothari, N. C. & Patel, M. H. (1993). *Phys. Rev. Lett.* **70**, 1112–1115.

Saldin, D. K., Harder, R., Shneerson, V. L., Vogler, H. & Moritz, W. (2000). *Theory and Computation for Synchrotron Radiation Spectroscopy*, edited by M. Benfatto, C. R. Natoli & E. Pace. *AIP Conference Proceedings*, No. 514, pp. 130–139. New York: American Institute of Physics.

Saldin, D. K., Shneerson, V. L. & Wild, D. L. (1997). *J. Imag. Sci. Technol.* **41**, 482–487.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

Sim, G. A. (1960). *Acta Cryst.* **13**, 511–512.

Sivia, D. S. (1996). *Data Analysis: a Bayesian Tutorial.* Oxford University Press.

Skarzynski, T. (1992). *J. Mol. Biol.* **224**, 671–683.

Somoza, J. R., Szöke, H., Goodman, D. M., Béran, P., Truckses, D., Kim, S.-H., & Szöke, A. (1995). *Acta Cryst.* A**51**, 691–708.

Szöke, A. (1993). *Acta Cryst.* A**49**, 853–866.

Szöke, A. (1998). *Acta Cryst.* A**54**, 543–562.

Szöke, A., Szöke, H. & Somoza, J. R. (1997). *Acta Cryst.* A**53**, 291–313.

Wilkins, S. W. (1983*a*). *Acta Cryst.* A**39**, 47–60.

Wilkins, S. W. (1983*b*). *Acta Cryst.* A**39**, 892–896.

Xiang, S., Carter, C. W., Bricogne, G. & Gilmore, C. J. (1993). *Acta Cryst.* D**49**, 193–212.