Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 18 March 2009
Accepted 19 May 2009

Determination of molecular envelopes from solvent contrast variation data

Victor Lo,[a] Richard L. Kingston[b] and R. P. Millane[a]*

[a]Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand, and [b]School of Biological Sciences, The University of Auckland, Auckland, New Zealand. Correspondence e-mail: rick.millane@canterbury.ac.nz

An algorithm is described for determining macromolecular envelopes from crystal diffraction amplitudes measured from a solvent contrast variation series. The method uses solvent contrast variation data that have been preprocessed to represent the structure-factor amplitudes of the envelope. The amplitudes are phased using an iterative projection algorithm that incorporates connectivity and compactness constraints on the envelope. The algorithm is tested by simulation on two protein envelopes and shown to be effective even in the absence of the very low resolution data, which are difficult to access experimentally.

1. Introduction

Despite recent advances, determination of the structures of large and complex macromolecular structures from crystal X-ray diffraction data can sometimes be problematic, as a result of the experimental difficulties of obtaining sufficiently accurate initial phase information. Solvent flattening, histogram matching and noncrystallographic symmetry averaging can be helpful in such cases (Kleywegt & Read, 1997; Podjarny *et al.*, 1996) by providing additional structural constraints, and in some cases they offer the potential of *ab initio* phasing (Lawrence, 1991; Rossmann, 1995). Application of these constraints requires knowledge of the region occupied by the molecule, the so-called molecular envelope. However, the molecular envelope is usually determined from preliminary electron density functions, calculated using experimentally derived phases, and so *ab initio* envelope determination presents a catch-22 situation. Solution scattering (Svergun & Stuhrmann, 1991; Chacon *et al.*, 2000; Svergun *et al.*, 2001; Hao, 2006; Svergun, 2007; Putnam *et al.*, 2007; Stuhrmann, 2008) or electron microscopy (Dodson, 2001; Hao, 2006; Navaza, 2008; Xiong, 2008) can be used to derive molecular envelopes; however, it would be useful if the molecular envelope could be determined directly from the structure-factor amplitudes of the crystal. One approach that potentially allows this is the method of contrast variation, which can be used to obtain estimates of the amplitudes that would be diffracted by the molecular envelope itself. The problem then reduces to phasing these derived amplitudes. In this paper we present a new approach for phasing of envelope structure-factor amplitudes.

Solvent contrast variation involves the collection and analysis of diffraction data from macromolecular crystals where the scattering contribution from the bulk solvent has been systematically varied. The potential uses of such information have long been known. By manipulating the electron density of the solvent, Bragg & Perutz (1952) were able to observe systematic changes in the intensity of the low-order diffraction data collected from haemoglobin crystals and infer the approximate dimensions of the molecule. A number of contrast variation experiments have subsequently been used to estimate molecular envelopes, usually by changing the salt or the salt concentration (Carter *et al.*, 1990). Another way of modulating diffraction from the bulk solvent is to disperse anomalous scatterers in it and make diffraction measurements about an absorption edge (Bricogne, 1993; Fourme *et al.*, 1995; Shepard *et al.*, 2000). An advantage of the latter approach is that isomorphism is conserved. Whichever method is used, the result, ideally, is the extraction of structure-factor amplitudes due to the molecular envelope alone, *i.e.* a function equal to unity within the envelope and zero outside. This function is sometimes referred to as the indicator function of the envelope, although we will refer to it here simply as the envelope.

Once estimates of the structure-factor amplitudes have been obtained, the problem is to phase these amplitudes to obtain the molecular envelope itself. As pointed out by Shepard *et al.* (2000), this phasing problem has a quite different character to the usual phase problems in crystallography. The corresponding electron density is not atomistic, it does not have the detailed structure of a low-resolution protein electron density, the electron density is far from being randomly distributed in the unit cell but is a rather compact binary function, and the number of (low-resolution) structure-factor amplitudes that are used as data is quite small. However, both Carter *et al.* (1990) and Fourme *et al.* (1995) argued that the problem has some similarities to small-molecule structure determination and used methods based on direct methods to phase the envelope diffraction amplitudes. Carter *et al.* (1990) used solvent contrast variation data and direct methods phasing to determine an 18 Å-resolution envelope of tryptophanyl-tRNA synthase from *Bacillus stearothermophilus*. Results were promising; however, these authors had the advantage of sixfold noncrystallographic symmetry at low resolution, and the method required considerable manual intervention. Fourme *et al.* (1995) showed that measurable anomalous scattering solvent contrast measurements could be made for two proteins, although there were experimental difficulties and the data were not used for envelope determination. They noted that the potential of the method for complex structures depends critically on the initial phase determination of the envelope amplitudes by direct methods, which has not yet been convincingly demonstrated. Shepard *et al.* (2000) took a different approach; they represented the envelope as a surface in spherical polar coordinates and parameterized the surface using spherical harmonics and a small number of coefficients. The coefficients are determined from the envelope structure-factor amplitudes using a nonlinear least-squares minimization procedure. Encouraging results were obtained using simulated data, although the authors noted that their method cannot represent general envelopes (since the actual surface function may be multi-valued), and there were difficulties with scaling the data and robustness of the gradient-based minimization procedure.

Neutron diffraction has also been explored for envelope determination by using differing H/D contents to vary the solvent scattering. Badger (1996) used solvent contrast neutron diffraction data from cubic insulin crystals, and application of a search procedure with a cost function that favours a binary histogram, to estimate the molecular envelope. However, the method is suitable only for the centric reflections and the search procedure is not feasible for a large data set.

Here we present an alternative method for determining molecular envelopes from the structure-factor amplitudes derived from solvent contrast variation. Our method is based on a recent study of properties of, and reconstruction algorithms for, the generic problem of reconstructing a compact, binary image from limited Fourier amplitude data (Lo & Millane, 2008). We showed there that the characteristics of molecular envelopes should allow a unique reconstruction from the structure-factor amplitudes alone. The basis of the reconstruction method is a set of constraints that embody the salient properties of molecular envelopes and a global optimization procedure based on the method of alternating projections. In the next section we briefly review contrast variation methods for deriving molecular envelope structure-factor amplitudes. In §3 we describe our algorithm and in §4 results of simulations for two protein crystals are presented. Concluding remarks are made in §5.

2. Envelope structure-factor amplitudes from solvent contrast variation data

The use of either solvents with different electron densities or solvents containing anomalous scatterers to derive the structure-factor amplitudes of the molecular envelope has been described previously (Carter *et al.*, 1990; Fourme *et al.*, 1995). The key elements of these calculations are briefly outlined here for the benefit of the reader.

Consider a unit cell with a protein molecule surrounded by solvent with electron density ρ_s. Let the envelope function, as defined in §1, be denoted $g(\mathbf{y})$, where $\mathbf{y}$ is the position in real space. The electron density in the unit cell, $f(\mathbf{y})$, can then be written as

$$f(\mathbf{y}) = \rho(\mathbf{y}) + \rho_s[1 - g(\mathbf{y})], \qquad (1)$$

where $\rho(\mathbf{y})$ is the electron density of the protein alone. The structure factor for $\mathbf{h} \neq \mathbf{0}$ is then

$$F_{\mathbf{h}} = F_{\mathbf{h}}^{\mathrm{P}} - \rho_s G_{\mathbf{h}}, \quad \mathbf{h} \neq \mathbf{0}, \qquad (2)$$

where $F_{\mathbf{h}}^{\mathrm{P}}$ is the structure factor of the protein and $G_{\mathbf{h}}$ is the structure factor of the envelope function. The equation for $\mathbf{h} = \mathbf{0}$ is somewhat different but is of little significance since $F_{\mathbf{0}}$ cannot be measured. Straightforward manipulation of equation (2) shows that the measured amplitudes are given by

$$|F_{\mathbf{h}}|^2 = |F_{\mathbf{h}}^{\mathrm{P}}|^2 + \rho_s^2 |G_{\mathbf{h}}|^2 - 2\rho_s \mathrm{Re}[F_{\mathbf{h}}^{\mathrm{P}} G_{\mathbf{h}}^*], \quad \mathbf{h} \neq \mathbf{0}, \qquad (3)$$

where Re[·] denotes the real part and * denotes complex conjugation. Equation (3) is linear in the three unknowns $|F_{\mathbf{h}}^{\mathrm{P}}|^2$, $|G_{\mathbf{h}}|^2$ and $\mathrm{Re}[F_{\mathbf{h}}^{\mathrm{P}} G_{\mathbf{h}}^*]$, so if data are collected for three different solvent electron densities ρ_s, then the three corresponding equations can be solved for these unknowns. In particular, the structure-factor amplitudes of the molecular envelope, $|G_{\mathbf{h}}|$, can be obtained. In practice the three data sets need to be put onto a common scale. In addition, some means of making the boundary between protein and solvent less step-like must be introduced. However, the description above shows the essence of the technique.

An alternative means of manipulating the scattering from the bulk solvent is to incorporate anomalous scatterers in the solvent and make measurements at different wavelengths. Advantages of this approach are that only a single crystal is required and there is no lack of isomorphism. The structure factors at wavelength λ are then given by

$$F_{\mathbf{h}}(\lambda) = F_{\mathbf{h}}^{\mathrm{P}} - [\rho_s + aK(\lambda)]G_{\mathbf{h}}, \quad \mathbf{h} \neq \mathbf{0}, \qquad (4)$$

where $K(\lambda)$ is the known, complex, wavelength-dependent scattering by the anomalous scatterers and a is a constant related to the concentration of anomalous scatterers in the solvent. Manipulation of equation (4) shows that the measured amplitudes are given by

$$\begin{aligned} |F_{\mathbf{h}}(\lambda)|^2 = {} & |F_{\mathbf{h}}^{\mathrm{P}}|^2 + |\rho_s + aK(\lambda)|^2 |G_{\mathbf{h}}|^2 \\ & - 2\mathrm{Re}[F_{\mathbf{h}}^{\mathrm{P}} \{\rho_s + aK(\lambda)\}^* G_{\mathbf{h}}^*], \quad \mathbf{h} \neq \mathbf{0}. \end{aligned} \qquad (5)$$

Hence, similarly to the previous case, measurement of $|F_{\mathbf{h}}(\lambda)|$ for different wavelengths (but fixed ρ_s) gives a system of linear equations that can be solved for $|G_{\mathbf{h}}|$. Since $K(\lambda)$ is complex, and so $F_{\mathbf{h}}(\lambda) \neq F_{-\mathbf{h}}(\lambda)$, two equations are obtained for each wavelength, and data for two wavelengths are in principal sufficient to solve for $|G_{\mathbf{h}}|$. In practice, the methods of multiple anomalous dispersion (Hendrickson, 1991) could be used to obtain a stable solution of equation (5).

3. Algorithm

The problem at hand is to reconstruct the molecular envelope $g(\mathbf{y})$ from its structure-factor amplitudes $|G_{\mathbf{h}}|$. The molecular envelope is a low-resolution object and so only the low-resolution amplitudes, say less than 7–10 Å resolution, are pertinent. However, an important practical consideration is that the amplitude data will be available only down to a minimum resolution of, say, 50 Å. The lower resolution limit presents a particular difficulty in this problem. We therefore consider amplitude data $|G_{\mathbf{h}}|$ between resolutions $d_{\min}$ and $d_{\max}$. We assume that such data $|G_{\mathbf{h}}|$ have been obtained, subject to the usual errors, from some form of solvent contrast variation experiment as described above.

The problem of reconstructing $g(\mathbf{y})$ corresponds to a usual macromolecular crystallographic phase problem, although the characteristics of $g(\mathbf{y})$ are different from those of a protein electron density. The envelope is a binary function, or image, and it forms a single, compact, connected object. We have studied this generic phase problem and shown that the binary constraint is sufficiently restrictive to uniquely define the envelope if all the low-resolution amplitudes are known, even in the presence of noise (Lo & Millane, 2008). Compactness and connectivity constraints are expected to compensate for the absence of some of the very low resolution amplitude data. Reconstruction of such an image should therefore be possible and we proposed an iterative projection algorithm to effect the reconstruction (Lo & Millane, 2008). Here we study the application of this algorithm to the recovery of molecular envelopes.

The information available to effect the reconstruction consists of the amplitudes $|G_{\mathbf{h}}|$ and several properties of the envelope, such that it is binary and that it forms a connected domain. The approach we take is to treat the problem as a constraint satisfaction problem and use the method of iterated projections (Elser, 2003; Millane, 2003). In this approach, an algorithm is used to find an image (in this case the envelope) that satisfies the available information, which is expressed as two sets of constraints. In the case at hand, the two constraints are the structure-factor amplitude data and the properties (binary, compactness, connectedness) of the envelope. The algorithm is iterative, and at each iteration a new 'iterate' is formed from the previous iterate by a combination of 'projections'. A projection consists of making the minimal change (in the squared deviation sense) to the iterate, such that it conforms to one of the constraints. For example, projecting an electron-density iterate onto the structure-factor amplitude constraint consists of calculating the structure factors, replacing the amplitudes by the measured amplitudes and then calculating a new electron density. The relationship between projections and many of the steps used in conventional electron-density modification is obvious. We note that, depending on the particular algorithm, an iterate is not necessarily an estimate of the solution, but that an estimate can be obtained by projecting an iterate onto any of the constraints. A variety of different iterative projection algorithms exist, which vary in the way that the projections are combined at each iteration (Millane, 2003; Marchesini, 2007). Here we use the 'difference map' (DM) algorithm (Elser, 2003), since this algorithm accommodates general constraints and is effective in avoiding stalling at near-solutions. Note that this is unrelated to the usual 'difference Fourier map' in crystallography.

It is convenient to formulate iterative projection algorithms as operations on points $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ in an N-dimensional abstract vector space R^N. A point $\mathbf{x}$ in this vector space represents a particular image (envelope) in the discretized unit cell, where N is the number of grid points in the unit cell, and the value of a component of $\mathbf{x}$, x_j, is the value of the image at the corresponding grid point $\mathbf{y}_j$, *i.e.* $x_j = g(\mathbf{y}_j)$. The projection of $\mathbf{x}$ onto a constraint A is denoted $P_A\,\mathbf{x}$ and is defined by

$$P_A\,\mathbf{x} = \operatorname*{argmin}_{\mathbf{x}' \in A} \|\mathbf{x}' - \mathbf{x}\|, \tag{6}$$

where A is a subset of R^N containing all images that satisfy the constraint, $\|\cdot\|$ is the Euclidean norm and $\operatorname{argmin}_{\mathbf{x}}[\alpha(\mathbf{x})]$ denotes the value of $\mathbf{x}$ that minimizes $\alpha(\mathbf{x})$. We denote the real-space constraint set by A and the structure-factor amplitude data constraint set by B. With this formalism, one iteration of the DM algorithm is defined by (Elser, 2003)

$$\begin{aligned}\mathbf{x}_{n+1} = \mathbf{x}_n + \beta\{&P_A[(1 + 1/\beta)P_B\mathbf{x} - (1/\beta)\mathbf{x}] \\ &- P_B[(1 - 1/\beta)P_A\mathbf{x} + (1/\beta)\mathbf{x}]\},\end{aligned} \tag{7}$$

where β is a constant such that $-1 \le \beta \le 1$ and $\mathbf{x}_n$ denotes the nth iterate. We note that equation (7) is the most commonly used form of the general DM algorithm (Elser, 2003). Note also that interchanging P_A and P_B is equivalent to changing the sign of β. The algorithm is started with a random (or otherwise) initial image $\mathbf{x}_0$ and convergence is monitored as the distance of the iterate from the constraints.

As noted previously (Elser, 2003; Lo & Millane, 2008), the iterate $\mathbf{x}$ itself of the DM algorithm is not an estimate of the solution. An estimate of the solution can be obtained by projecting the iterate onto one of the constraints. Here we use an estimate of the envelope, denoted $\hat{\mathbf{g}}_n$, from an iterate $\mathbf{x}_n$ computed as

$$\hat{\mathbf{g}}_n = P_A[(1 + 1/\beta)P_B\mathbf{x}_n - (1/\beta)\mathbf{x}_n]. \tag{8}$$

Since this quantity is calculated at each iteration anyway, no additional computational cost is incurred. We note that, for $\beta = 1$, this corresponds to computing a '$2F_{\rm o} - F_{\rm c}$' map.

The Fourier amplitude projection P_B is given by

$$P_B\,\mathbf{x} = F^{-1}[\tilde{P}_B F[\mathbf{x}]], \tag{9}$$

where $F[\cdot]$ and $F^{-1}[\cdot]$ denote the Fourier and inverse Fourier transforms, respectively, and $\tilde{P}_B$ is the Fourier amplitude projection in Fourier space given by

$$\tilde{P}_B X_{\mathbf{h}} = \begin{cases} sM_{\mathbf{h}} \exp\{i\varphi(X_{\mathbf{h}})\} & \text{if } \mathbf{h} \in Q \\ X_{\mathbf{h}} & \text{if } \mathbf{h} \notin Q, \end{cases} \tag{10}$$

where $i = (-1)^{1/2}$, $\{X_{\mathbf{h}}\} = F[\mathbf{x}]$, $\varphi(\cdot)$ denotes the phase, $M_{\mathbf{h}}$ denotes the measured structure-factor amplitudes of the envelope derived from the measured data, s is a scale factor and Q denotes the set of reciprocal lattice points where the data are measured (*i.e.* between the resolutions $d_{\min}$ and $d_{\max}$). Simulations showed that the method used to estimate the scale factor is important and this is discussed further in §4.

The projection P_A corresponds to adjusting an iterate such that it conforms to our knowledge of molecular envelopes. We therefore adjust an iterate such that (1) it is a binary function, (2) it has the correct solvent-excluded volume (this can generally be estimated from the protein molecular weight, the space group and the unit-cell dimensions) and (3) it forms a single connected domain with no 'holes'. The no-holes constraint is not restrictive and can be relaxed if required. An exact projection onto these constraints is not feasible, and the approximate projection P_A that we use is described in detail by Lo & Millane (2008) and consists of the following steps.

The first step is projection onto the set of binary images with the correct solvent-excluded volume. The fractional solvent-excluded volume is denoted f. This projection is denoted P_{BF}, and involves setting the fN largest values of $\mathbf{x}$ to 1 and the remainder 0, *i.e.*

$$P_{BF}\,x_j = \begin{cases} 0 & \text{if } x_j \notin S(f) \\ 1 & \text{if } x_j \in S(f), \end{cases} \tag{11}$$

where $S(f)$ is the set of the largest fN values of $\mathbf{x}$.

In the second step, connected binary regions, or objects, that are larger than a threshold size, denoted l, are retained and the remaining objects deleted. This operation is denoted P_C and is defined by

$$P_C\,x_j = \begin{cases} 1 & \text{if } j \in L(l) \\ 0 & \text{if } j \notin L(l), \end{cases} \tag{12}$$

where $L(l)$ denotes the set of grid points that belong to objects of size (number of grid points) greater than l. The threshold l is given by $l = \alpha f N$, with the constant $\alpha \simeq 0.1$ (Lo & Millane, 2008). The effect of the operation P_C is to favour a single object as the iterations proceed.

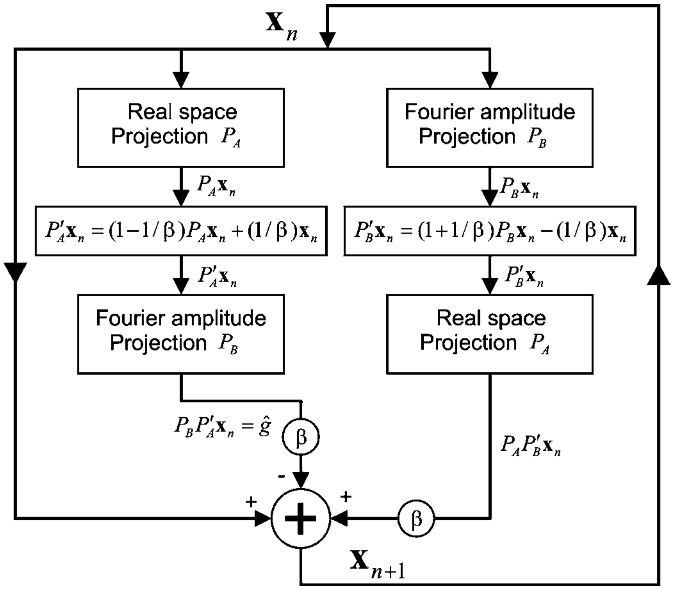
$\mathbf{x}_n$
Real space Projection P_A
Fourier amplitude Projection P_B
$P_A\mathbf{x}_n$
$P_B\mathbf{x}_n$
$P'_A\mathbf{x}_n = (1-1/\beta)P_A\mathbf{x}_n + (1/\beta)\mathbf{x}_n$
$P'_B\mathbf{x}_n = (1+1/\beta)P_B\mathbf{x}_n - (1/\beta)\mathbf{x}_n$
$P'_A\mathbf{x}_n$
$P'_B\mathbf{x}_n$
Fourier amplitude Projection P_B
Real space Projection P_A
$P_B P'_A\mathbf{x}_n = \hat{g}$
β
$P_A P'_B\mathbf{x}_n$
-
+
+
β
$\mathbf{x}_{n+1}$

Figure 1
Block diagram of the algorithm.

Table 1
Parameters of the proteins A and B.

Protein	Cell dimensions (Å)	Number of grid points	Grid spacing (Å)	f
A	77.2 × 176.7 × 51.1	18 × 40 × 12	4.3	0.35
B	54.2 × 65.3 × 73.6	18 × 20 × 24	3.1	0.57

In the third step, solvent volumes larger than a threshold size are retained and smaller 'holes' removed. This operation is denoted P_{SC} and is identical to the previous step applied to the negative of the image, so that

$$P_{SC}\,\mathbf{x} = 1 - P_C(1 - \mathbf{x}). \tag{13}$$

The full real-space projection, denoted P_A, consists of concatenation of the above three projections, *i.e.*

$$P_A\,\mathbf{x} = P_{SC} P_C P_{BF}\,\mathbf{x}. \tag{14}$$

A block diagram of the algorithm used is shown in Fig. 1. If the solvent-excluded volume f is close to 0.5, the algorithm may converge to the negative solution (solvent and protein regions interchanged). This solution can be avoided by a small change to the algorithm, which involves checking the negative solution as described by Lo & Millane (2008).

4. Simulations

The algorithm was tested by simulation on two molecular envelopes derived from solved protein structures taken from the Protein Data Bank. The two proteins are the alkaline protease from *Pseudomonas aeruginosa* (Miyatake *et al.*, 1995), and human galectin-7 (Leonidas *et al.*, 1998). For convenience, we refer to these two proteins as A and B, respectively. Both structures have space group $P2_12_12_1$ (the asymmetric unit is 1/4 of the unit cell). The unit-cell dimensions, sampling grid size and grid spacings are listed in Table 1. The solvent-excluded volumes for the two proteins are quite different, with $f = 0.35$ for protein A and $f = 0.57$ for protein B.

Molecular envelopes were determined from each atomic model using standard procedures (Wang, 1985; Leslie, 1987), as implemented in the program *DM* (Cowtan, 1994). The averaging radius for envelope generation was 8 Å. The Fourier amplitudes of the envelope were calculated by the discrete Fourier transform, a scale factor was applied, 5% r.m.s. Gaussian noise was added, and the amplitudes within a resolution shell between 40 and 7 Å were used as data for image reconstruction. The algorithm was started with a random binary image. Although the algorithm does not break any crystallographic symmetry present, the $P2_12_12_1$ crystallographic symmetry is maintained by averaging the image over the four asymmetric units at the end of each iteration to prevent rounding errors from accumulating.

In practice, the scale factor s [equation (10)] needs to be determined for application of the method. In some cases the

Table 2
Results for the two protein envelopes.

Protein	Resolution range (Å)	Runs	Converged runs	Correct solutions	Incorrect solutions
A	40–7	5	2	2	0
A	50–7	5	5	5	0
B	40–7	5	5	1	4
B	50–7	5	5	5	0

scale factor may be estimated from crystal packing data and the low-resolution diffraction amplitudes, but the problem is not well determined if the available low-resolution diffraction data are limited. It was found that determination of s was crucial and not straightforward. Our solution to this problem is described here. An obvious approach is, starting with an estimate, to refine the scale factor as part of the iterative reconstruction procedure by scaling the mean estimated structure-factor amplitudes from the envelope to the mean of the measured amplitudes. The difficulty, however, as mentioned above, is that the low-resolution amplitudes, which have the greatest effect on the scale factor, are not available. Although this approach was successful in some cases, it was not successful in general. This is because, since we are solving the *ab initio* problem, the early estimates of the envelope are highly incorrect, which can lead to a poor initial estimate of the scale factor. This incorrect scale factor may trap the iterations into a local minimum from which it is difficult to escape. The following method of obtaining a good estimate of the scale factor was found to be effective.

The asymmetric unit is chosen to minimize its aspect ratio, *i.e.* to be as close to cubic as possible, and an ellipsoid is placed at the centre of each asymmetric unit. The ratio of the semi-axes of the ellipsoid is the same as the ratio of the axes of the asymmetric unit, and the size of the ellipsoid is chosen such that the total volume of the ellipsoids is equal to the known volume of the envelope. For crystals of low solvent content the ellipsoids may be truncated by the faces of the asymmetric unit, and the size of the ellipsoid would then need to be increased. This method was found to be suitable in all cases we studied. For the orthorhombic case $P2_12_12_1$ considered here, if the cell dimensions satisfy $a > b > c$, there are four asymmetric units of dimensions $a/2 \times b/2 \times c$. The scale factor is then estimated as

$$s = \sum_{\mathbf{h}\in Q'} |W_\mathbf{h}|^2 \Big/ \sum_{\mathbf{h}\in Q'} M_\mathbf{h}^2, \qquad (15)$$

where $|W_\mathbf{h}|$ are the structure-factor amplitudes of the ellipsoid model and Q' is an appropriate resolution range. It was found that a resolution range of 25–7 Å was suitable. Simulations showed that this method generally gave scale factors within 5% of the correct value. However, even starting with this value and refining the scale factor as described above sometimes led to divergence. Therefore, it was found to be most satisfactory to lock the scale factor at the estimated value. A locked scale factor with an error of up to 5% did not significantly affect convergence or the quality of the solution.

Since the estimate of the envelope $\hat{\mathbf{g}}_n$ [equation (8)] satisfies the real-space constraints, convergence of the algorithm was monitored by calculating the quadratic R factor as

$$R_n'' = \sum_{\mathbf{h}\in Q} (|\hat{G}_{\mathbf{h},n}| - sM_\mathbf{h})^2 \Big/ \sum_{\mathbf{h}\in Q} s^2 M_\mathbf{h}^2, \qquad (16)$$

where $|\hat{G}_{\mathbf{h},n}|$ are the structure-factor amplitudes of $\hat{\mathbf{g}}_n$. The error in the envelope given by

$$T_n = ||\mathbf{g} - \hat{\mathbf{g}}_n||^2 / ||\mathbf{g}||^2, \qquad (17)$$

where $\mathbf{g}$ is the true envelope, was also calculated to monitor the accuracy of the solution. The proportion of grid points in error is then equal to fT_n. Clearly, this metric can only be calculated for simulations where the original envelope is known. It was found that $T_n < 0.2$ corresponds to a good estimate of the true envelope.

In general, as the algorithm proceeds the iterates wander around the solution space before finding a solution and closing in with a distinctive fall in the error metric (Lo & Millane, 2008). Because of the noise, there may be no solution that exactly satisfies all the constraints, and the nature of the DM

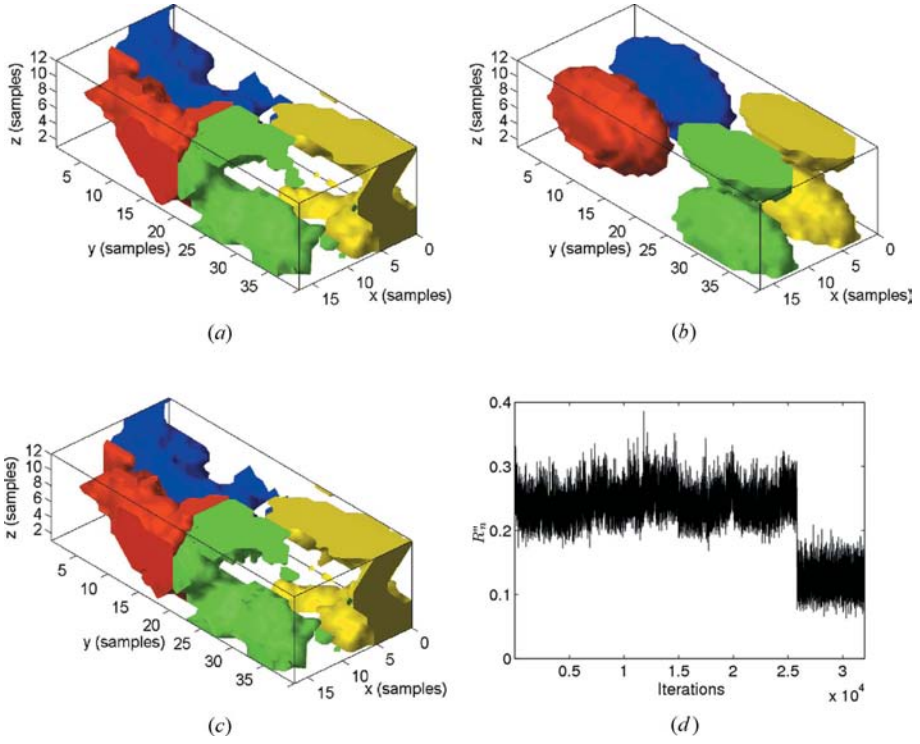

Figure 2
(*a*) The true envelope, (*b*) the ellipsoid model, (*c*) the final solution and (*d*) R_n'' *versus* iteration, for protein *A*. Symmetry-equivalent regions in the unit cell are represented by different colours to aid interpretation.

algorithm is such that the iterates will then move away from the near solution. Therefore, a large number of iterations are used, and the solution with the minimum error metric R''_n is chosen. In practice, it is best to run the algorithm a few times with different starting envelopes and select the solution with the best agreement index R''_n.

The DM algorithm was run with $\beta = 0.9$ for 1.5×10^5 iterations with 5% noise on the data for envelopes A and B. Values $0.7 < |\beta| < 1.0$ worked well; however, positive values of β gave slightly faster convergence than negative values. The scale factor was estimated as described above. Five runs were made using different random starting envelopes for each resolution range. The results are summarized in Table 2. The table shows the resolution range used, the total number of runs, the number of converged runs ($R''_n < 0.1$), the number of successful converged runs ($T_n < 0.2$) and the number of incorrect solutions obtained.

For protein A with data in the range 40–7 Å, two of the five runs converged. For all of the converged runs, an accurate reconstruction of the envelopes was obtained, with T_n in the range 0.02–0.05, *i.e.* no incorrect solutions were obtained. If the lower resolution limit is reduced to 50 Å, all runs converged to the correct solution. The results for one of the converged runs for which $R''_n = 0.056$ and $T_n = 0.033$ are shown in Fig. 2. The plot of R''_n *versus* iteration shows a sharp drop at about iteration 2500, followed by erratic movement of the iterate around the correct solution. The true envelope, the ellipsoid model and the reconstructed envelope are also shown in the figure. The reconstructed envelope is seen to be a good estimate of the true envelope. The algorithm has therefore been successful in this case.

For protein B with data in the range 40–7 Å, all five of the runs converged. Of these, one gave the correct solution ($T_n < 0.2$) and four gave incorrect solutions ($T_n > 0.2$). Therefore, although convergence could be obtained, the existence of multiple solutions that replicate the data indicates that, in this case, the data are insufficient to define a unique solution. Extending the lower resolution limit down to 50 Å, all five runs converged, and all gave the correct solution ($T_n < 0.2$). In this case, therefore, more low-resolution diffraction data are needed to uniquely define the envelope. The results for one of the converged runs for which $R''_n = 0.018$ and $T_n = 0.09$ are shown in Fig. 3. In this case the error R''_n drops quite rapidly and the algorithm is stable at the solution. The true envelope, the ellipsoid model and the reconstructed envelope are also shown in the figure. Although $T_n = 0.09$, for the reconstructed envelope, the reconstruction is quite accurate with only 5% of the grid points misclassified.

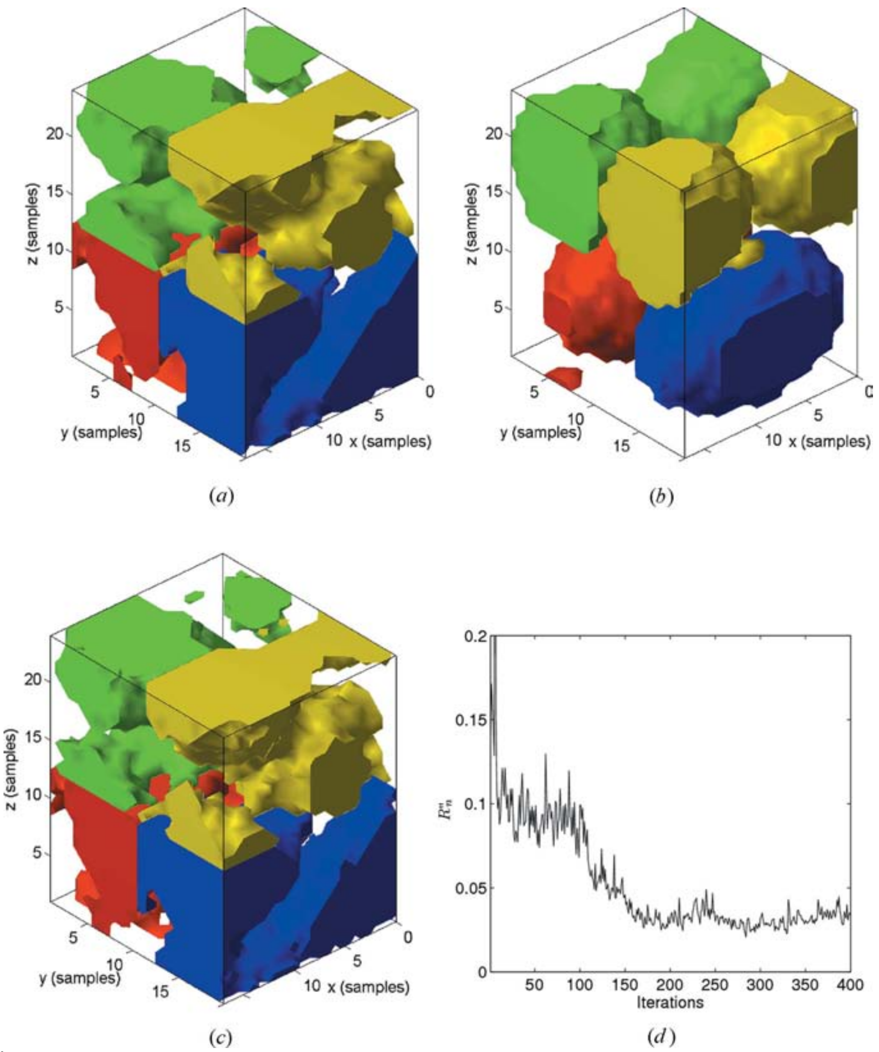

Figure 3
(*a*) The true envelope, (*b*) the ellipsoid model, (*c*) the final solution and (*d*) R''_n *versus* iteration, for protein *B*.

5. Conclusions

The structure-factor amplitudes of a molecular envelope obtained from solvent contrast variation experiments, when coupled with *a priori* information on envelopes, can uniquely define the envelope. Incorporation of connectivity and compactness constraints into an iterative projection algorithm gives an effective way of reconstructing envelopes from such data. Simulations with real envelopes and realistic levels of noise and missing data indicate that this algorithm may be practical. Advantages of the algorithm are that it is automatic and requires no additional information. The solution to the problem is sensitive to missing low-resolution data and to an accurate determination of the scale factor.

VL is the recipient of a New Zealand Tertiary Education Commission Top Achiever Doctoral Scholarship.

References

Badger, J. (1996). *Basic Life Sci.* **64**, 333–343.
Bragg, W. L. & Perutz, M. F. (1952). *Acta Cryst.* **5**, 277–283.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Carter, C. W., Crumley, K. V., Coleman, D. E., Hage, F. & Bricogne, G. (1990). *Acta Cryst.* A**46**, 57–68.
Chacon, P., Diaz, J. F., Moran, F. & Andreu, J. M. (2000). *J. Mol. Biol.* **299**, 1289–1302.
Cowtan, K. (1994). *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*, No. 31, pp. 34–38.
Dodson, E. J. (2001). *Acta Cryst.* D**57**, 1405–1409.
Elser, V. (2003). *J. Opt. Soc. Am. A*, **20**, 40–55.
Fourme, R., Shepard, W., Kahn, R., l'Hermite, G. & Li de La Sierra, I. (1995). *J. Synchrotron Rad.* **2**, 36–48.
Hao, Q. (2006). *Acta Cryst.* D**62**, 909–914.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Kleywegt, G. J. & Read, R. J. (1997). *Structure*, **5**, 1557–1569.
Lawrence, M. C. (1991). *Q. Rev. Biophys.* **24**, 399–424.
Leonidas, D. D., Vatzaki, E. H., Vorum, H., Celis, J. E., Madsen, P. & Acharya, K. R. (1998). *Biochemistry*, **37**, 13930–13940.
Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.
Lo, V. L. & Millane, R. P. (2008). *J. Opt. Soc. Am. A*, **25**, 2600–2607.
Marchesini, S. (2007). *Rev. Sci. Instrum.* **78**, 1–10.
Millane, R. P. (2003). *Proc. Oceans 2003, CD-ROM, IEEE*, pp. 2714–2719.
Miyatake, H., Hata, Y., Fujii, T., Hamada, K., Morihara, K. & Katsube, Y. (1995). *J. Biochem.* **118**, 474–479.
Navaza, J. (2008). *Acta Cryst.* D**64**, 70–75.
Podjarny, A. D., Rees, B. & Urzhumtsev, A. G. (1996). *Methods Mol. Biol.* **56**, 205–226.
Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007). *Q. Rev. Biophys.* **40**, 191–285.
Rossmann, M. G. (1995). *Curr. Opin. Struct. Biol.* **5**, 650–655.
Shepard, W., Kahn, R., Ramin, M. & Fourme, R. (2000). *Acta Cryst.* D**56**, 1288–1303.
Stuhrmann, H. B. (2008). *Acta Cryst.* A**64**, 181–191.
Svergun, D. I. (2007). *J. Appl. Cryst.* **40**, s10–s17.
Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.
Svergun, D. I. & Stuhrmann, H. B. (1991). *Acta Cryst.* A**47**, 736–744.
Wang, B. C. (1985). *Methods Enzymol. B*, **115**, 90–112.
Xiong, Y. (2008). *Acta Cryst.* D**64**, 76–82.