Acta Cryst A
FOUNDATIONS
ADVANCES

ISSN 2053-2733

Iterative projection algorithms in protein crystallography. II. Application

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

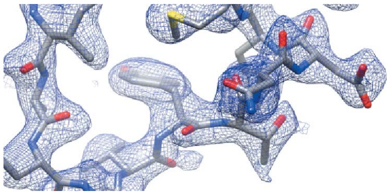
[a]Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand, and [b]School of Biological Sciences, The University of Auckland, Auckland, New Zealand. *Correspondence e-mail: rick.millane@canterbury.ac.nz

Received 23 December 2014
Accepted 18 March 2015

Edited by H. Schenk, University of Amsterdam, The Netherlands

Keywords: iterative projection algorithms; phasing; density modification; phase determination.

Iterative projection algorithms (IPAs) are a promising tool for protein crystallographic phase determination. Although related to traditional density-modification algorithms, IPAs have better convergence properties, and, as a result, can effectively overcome the phase problem given modest levels of structural redundancy. This is illustrated by applying IPAs to determine the electron densities of two protein crystals with fourfold non-crystallographic symmetry, starting with only the experimental diffraction amplitudes, a low-resolution molecular envelope and the position of the non-crystallographic axes. The algorithm returns electron densities that are sufficiently accurate for model building, allowing automated recovery of the known structures. This study indicates that IPAs should find routine application in protein crystallography, being capable of reconstructing electron densities starting with very little initial phase information.



1. Introduction

Electron-density modification is commonly used in protein crystallography to refine initial phase estimates by applying real-space constraints. Density-modification procedures (Cowtan, 2010; Terwilliger, 2003; Abrahams & Leslie, 1996) are integrated into the automated structure solution packages now in widespread use. Success, however, depends on the accuracy of initial phases, which are generally obtained experimentally using the methods of isomorphous replacement or anomalous dispersion, or computationally, using the method of molecular replacement. When initial phase estimates are poor, density modification may fail to determine high-resolution phases that are sufficiently accurate for model building. There is, therefore, interest in the continued development of methods for structure determination in the absence of reliable phase estimates.

We have shown previously that with fairly minimal structural redundancy, the electron density is uniquely determined by the structure-factor amplitudes alone (Millane & Lo, 2013). Furthermore, we also showed that it should be possible to determine protein electron densities from the diffraction amplitudes by using iterative projection algorithms that have better global convergence properties than conventional electron-density modification algorithms (Millane & Lo, 2013). In this paper, we demonstrate the effectiveness of these algorithms by using them to determine two tetrameric protein crystal structures starting with only the crystallographic diffraction amplitudes, a low-resolution envelope and the position of the non-crystallographic symmetry axes.

This paper is organized as follows. In §2 we briefly review the information needed for phasing in protein crystallography

http://dx.doi.org/10.1107/S2053273315005574

and the nature of iterative projection algorithms. Some specific details of the implementation of the iterative projection algorithm for protein crystallography are described in §3. Results of the application to the two proteins are described in §4. Concluding remarks are made in §5.

2. Background

Although the macromolecular crystallographic phase problem is generally underdetermined in the absence of additional experimental data, it is well known that additional real-space information, or structural redundancy, constrains the problem (Crowther, 1969; Bricogne, 1974; Millane, 1990; Liu *et al.*, 2012; Millane & Lo, 2013). If a low-resolution molecular envelope and the position of any non-crystallographic symmetry (NCS) operators are known, then a unique solution to the phase problem can be expected in the absence of any additional information if the order of the NCS, R, satisfies $R > 2f$, where f is the proportion of the unit cell occupied by protein (Millane & Lo, 2013). However in practice, in the presence of noise and missing data, and based on the results of Liu *et al.* (2012), a more realistic requirement is that the order of the NCS satisfies (Millane & Lo, 2013)

$$R > 3f. \tag{1}$$

Therefore, fairly minimal structural redundancy should be sufficient in practice to uniquely determine the electron density. With insufficient constraints the solution is non-unique and no phase retrieval algorithm will locate the correct solution. However, although with sufficient constraints the solution is unique, a phase retrieval algorithm may still fail to find the solution because the associated optimization problem is highly non-convex and location of the solution is nontrivial. In Millane & Lo (2013) we have proposed iterative projection algorithms as an effective method for finding the solution in the absence of initial phase information.

Iterative projection algorithms for phase determination in protein crystallography are described by Millane & Lo (2013), and here we briefly review the background of these algorithms for the purposes of the current paper. The reader is referred to Millane & Lo (2013) for more details. These algorithms utilize the same kinds of real-space constraints, such as solvent flatness, non-crystallographic symmetry, histograms, *etc.*, as are used in conventional electron-density modification algorithms. Constraint information is generally incorporated by adjusting the electron density in a minimal way such that the corresponding real-space constraint is satisfied. Such a step can be identified as a 'projection' onto the constraint (Bricogne, 1974; Millane, 1990; Marks *et al.*, 1999; Elser, 2003*a*; Millane & Lo, 2013). Incorporation of the diffraction data can be viewed in a similar way as a projection of the electron density onto a diffraction-amplitude constraint, *i.e.* the electron density is adjusted so that its Fourier amplitudes are equal to the measured structure-factor amplitudes.

The important difference between general iterative projection algorithms and conventional density-modification algorithms is the way in which the projections are incorporated (Millane & Lo, 2013). In conventional density modification, the estimated electron density is obtained by simply alternately projecting it onto the real-space and reciprocal-space constraints (Bricogne, 1974; Millane & Lo, 2013). In the image processing literature, this algorithm is often referred to as the 'error-reduction algorithm' (Fienup, 1982). The difficulty with this algorithm, however, is that it is prone to 'stagnation' if it is not started with phases that are reasonably close to their correct values. When the algorithm stagnates, or reaches a so-called *fixed point*, the electron density estimate returns to the same (incorrect) value at subsequent iterations, and so no progress is made towards the correct solution. The error-reduction algorithm, therefore, has poor global convergence properties, and this feature is known to be due to the non-convexity of the diffraction-amplitude constraint (Millane, 1990; Millane & Lo, 2013).

There is, however, a class of more general iterative projection algorithms that are more resistant to stagnation and have better global convergence properties (Elser, 2003*a*; Marchesini, 2007; Thumiger & Zanotti, 2009; Millane & Lo, 2013). The difference between these algorithms and conventional density modification is the way in which the projections are used at each iteration of the algorithm. For the purposes of describing these algorithms, the electron density is represented by an N-dimensional vector $\mathbf{x}$ whose components are the samples of the electron density at the N grid points in the unit cell or in the asymmetric unit. An iterative projection algorithm proceeds by updating an electron density 'iterate', denoted by the vector $\mathbf{x}_n$, at each iteration n, according to some update rule.

Conventional density modification, or the error-reduction algorithm, involves a simple alternation of projections of the electron density onto the real-space and reciprocal-space constraints, and so the update rule is

$$\mathbf{x}_{n+1} = P_A P_B \mathbf{x}_n, \tag{2}$$

where P_A and P_B denote the projections onto the real-space and reciprocal-space constraints, respectively. The more sophisticated iterative projection algorithms that have better global convergence properties use more complicated update rules than (2). We note that solvent flipping (Abrahams, 1997), and also charge flipping (Oszlanyi & Suto, 2008; Palatinus, 2013), are particular cases of more general iterative projection algorithms (Millane & Lo, 2013). Solvent flipping can speed the convergence of conventional density modification but it does not have a sufficiently large radius of convergence to successfully recover the phases if the initial phase estimates are poor. Charge flipping is more suitable for small-molecule phasing where atomic resolution data are available.

A number of iterative projection algorithms that have good global convergence properties have been described [see, for example, Marchesini (2007) and Millane & Lo (2013)]. These different algorithms tend to have similar convergence properties. One of the first such algorithms was the 'hybrid input–output' algorithm developed for applications in optics (Fienup, 1982). This algorithm has had some application in protein crystallography (Millane & Stroud, 1997; van der Plas

& Millane, 2000; Liu *et al.*, 2012). Here, we choose the 'difference map' algorithm of Elser (2003*a*) for our experiments. Note that the 'difference map' algorithm is unrelated to the difference Fourier synthesis routinely employed in protein crystallography (Henderson & Moffat, 1971). The update rule for the difference-map algorithm is

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \beta[P_A F_B(1/\beta)\mathbf{x}_n - P_B F_A(-1/\beta)\mathbf{x}_n], \qquad (3)$$

where $F_A(-1/\beta)$ and $F_B(1/\beta)$ are referred to as *relaxed projections* onto the real-space and reciprocal-space constraints, respectively, and are defined in terms of the projections P_A and P_B by

$$F_A(-1/\beta)\mathbf{x} = P_A x + (-1/\beta)(P_A x - x), \qquad (4)$$

and similarly for $P_B(1/\beta)$. The difference-map algorithm has the single parameter β and values $\beta \approx 0.7$ are usually suitable. The difference-map algorithm has been applied to *ab initio* phasing in small-molecule crystallography (Elser, 2003*b*).

A characteristic of these kinds of algorithms is that they tend to be unstable near a fixed point that does not correspond to a solution. This feature is related to their resistance to stagnation and their ability to explore the parameter space. Since, as a result of various errors and noise, the solution with experimental data will never be exact, after initially approaching the solution, the iterate will sometimes drift away from the solution. Measures may need to be taken to arrest this divergent behaviour.

It is important to note that the iterate $\mathbf{x}_n$ in an iterative projection algorithm is not itself generally an estimate of the electron density, but is an auxiliary function that is used by the algorithm to search the parameter space. Therefore, when the algorithm has converged, *i.e.* $\mathbf{x}_{n+1} \approx \mathbf{x}_n$, $\mathbf{x}_n$ does not generally represent an estimate of the solution. However, on convergence, an estimate of the solution can be obtained from the iterate. In the case of the difference-map algorithm, the solution, denoted $\hat{\mathbf{x}}$, is given by

$$\hat{\mathbf{x}} = P_A F_B(1/\beta)\mathbf{x}_n^*, \qquad (5)$$

where $\mathbf{x}_n^*$ denotes the iterate at convergence (Elser, 2003*a*; Millane & Lo, 2013).

In the context of protein crystallography, assuming that at least a molecular envelope is available, the difference-map algorithm can be applied by starting with a random electron density within the envelope, and applying the update rule (3) at each iteration using the projection operators P_A and P_B. Equivalently, one could initiate the algorithm in reciprocal space, rather than real space, by starting with random phases. The projection operators correspond to the usual density-modification steps used in conventional electron-density modification algorithms as described above (Millane & Lo, 2013). In the experiments described here, we use solvent flatness and NCS constraints in real space. The details are described in §3.3.

There has been some exploratory application of iterative projection algorithms in protein crystallography. Millane & Stroud (1997) and van der Plas & Millane (2000) adapted the hybrid input–output algorithm to incorporate an NCS constraint and applied it to reconstruction of an icosahedral virus with fivefold NCS, starting with a spherical shell and using simulated data. Good maps were obtained at 8 Å resolution. Lo *et al.* (2009) applied the difference-map algorithm to the determination of molecular envelopes in protein crystals from simulated solvent contrast variation data by incorporating binary constraints as well as connectivity/compactness constraints. Lo & Millane (2010) applied the difference-map algorithm to the reconstruction of an icosahedral virus starting from a spherical shell using experimental data and a fivefold NCS constraint, which gave a good reconstruction at about 3 Å resolution. Liu *et al.* (2012) applied the hybrid input–output algorithm with a solvent flatness constraint to a number of solved proteins with high (>65%) solvent content and data to between 2.0 and 2.8 Å resolution, starting with molecular envelopes calculated from the atomic coordinates. The algorithm was supplemented with conventional histogram matching and resulted in interpretable maps.

Inspection of (1) shows that in the absence of non-crystallographic symmetry (*i.e.* $R = 1$), a unique solution is expected only if the solvent content exceeds about 67%. However, this accounts for only about 7% of previously characterized protein crystals (Weichenberger & Rupp, 2014). Therefore, for most proteins, additional real-space constraint information will be needed in order to obtain a unique solution to the phase problem. In this paper we exploit NCS as the additional constraint. We apply the difference-map algorithm to phasing diffraction data from several protein crystals with lower solvent contents (~50%) starting with only a low-resolution envelope and employing solvent flatness and NCS constraints. The results show the potential of iterative projection algorithms for phasing protein structures with no initial phase information.

3. Methods

The basic reconstruction algorithm used here is the difference-map algorithm as described in §2, and a number of details of the implementation for protein crystallography are described in this section.

3.1. Data weighting and phase extension

In conventional electron-density modification, medium-resolution experimental phases are frequently extended to high resolution by incorporating the higher resolution amplitudes, stepwise in resolution shells and re-refining all the phases at each step (Lawrence, 1991; Rossmann, 1995). The effect of this stepwise incorporation of the higher resolution data is to improve the likelihood of correct phase determination at higher resolution compared with attempting to phase all the data in a single step. Our experiments with the application of iterative projection algorithms have shown that a similar form of phase extension in which the higher resolution amplitude data are gradually incorporated in a stepwise fashion improves the likelihood and speed of convergence.

It is well known that windowing (or apodizing or tapering) spectral data with a filter that has a smooth fall-off in Fourier space is preferable to an abrupt truncation, because it suppresses ringing in the other (real space) domain, and this technique is widely applied in signal processing, optics and imaging (Waser & Shomaker, 1953; Harris, 1978; Bracewell, 1986). We take this approach here, and rather than using a sharp resolution cutoff at each phase extension step, the diffraction-amplitude data are tapered with a Gaussian function. The resolution is extended in steps, and at the *m*th step the diffraction amplitudes are multiplied by a weight function, $w_m(s)$, given by

$$w_m(s) = \exp(-s^2/2\sigma_m^2), \tag{6}$$

where $1/s$ is the resolution in ångströms associated with each amplitude and the standard deviation σ_m determines the degree to which the high-resolution data are incorporated. In the experiments performed here, 11 resolution steps were used with weighting functions that have half-heights (*i.e.* where $s = 1.2\sigma$) at 30, 20, 10, 8, 6, 5, 4, 3, 2, 1 Å resolution, and at the final resolution step all the diffraction data are used with uniform weighting, as shown in Fig. 1. The reconstruction algorithm is run for a number of iterations at the lowest resolution step and then the value of σ_m is increased to the value for the next step and another set of iterations is run. The resolution is increased stepwise in this manner until the final step where no weighting is applied. As a result of the Gaussian weighting, the resolution is not readily defined at each step and so we sometimes denote the progression to higher resolution with the labels 1 through to 11 which refer to the corresponding weight functions shown in Fig. 1.

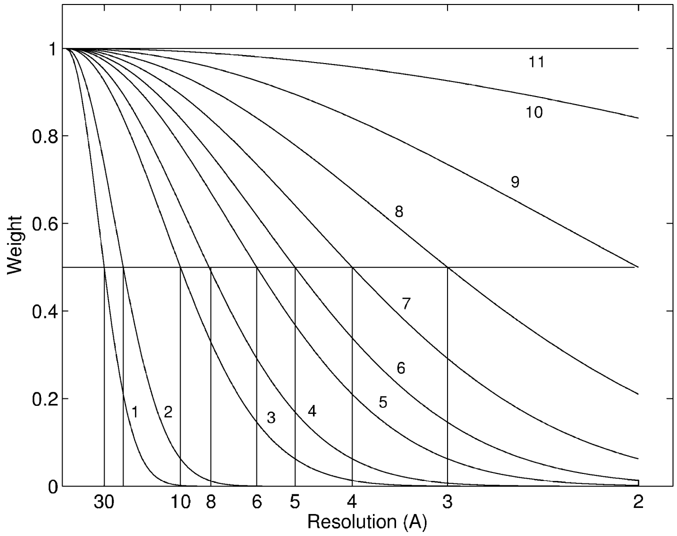


Figure 1
The weighting function $w_m(s)$ *versus* resolution ($1/s$) for each resolution step. The vertical lines show the resolution at the half-height of each weighting function.

3.2. Envelope position and orientation

We assume here that we are studying a symmetric protein oligomer; that a low-resolution estimate of the molecular envelope is available; and that this envelope can be correctly positioned and oriented within the unit cell. This appears to be an accessible starting point using existing methodology. If even a poor heavy atom derivative can be prepared, SAD/MAD or SIRAS phasing would generate the necessary information (McCoy & Read, 2010; Hendrickson, 2014). Alternatively, if a low-resolution envelope for the oligomer was generated experimentally using small-angle X-ray solution scattering (SAXS) or transmission electron microscopy (TEM), it might be positioned within the unit cell. There is a precedent for using low-resolution TEM-derived image reconstructions and SAXS-derived envelopes as search models in molecular replacement (Urzhumtsev & Podjarny, 1995; Dodson, 2001; Hao, 2006; Navaza, 2008; Xiong, 2008; Hong & Hao, 2009; Trapani *et al.*, 2010; Stuart & Abrescia, 2013). In many cases the orientation of the oligomer within the crystal can be deduced from inspection of the self-rotation function (Tong & Rossmann, 1997; Sawaya, 2007), providing an independent check on the validity of any solution (Dodson, 2001; Xiong, 2008).

For the purposes of our simulations, we generated molecular envelopes by Gaussian kernel smoothing the electron density calculated from the atomic coordinates (Gaussian half-height 10 Å, and applying a thresholding step to generate a binary envelope encompassing the correct fraction protein [see Wang (1985)]. For the simulations, we utilized crystallographic data for two tetramers with 222 point-group symmetry that crystallize in space group $P2_12_12_1$ with a tetramer in the asymmetric unit, giving rise to fourfold NCS, and the rotational NCS operations associated with each tetramer were determined from the atomic coordinates.

3.3. Projections

In the experiments described here, we use solvent flatness and NCS constraints to determine the projection operator (P_A) in real space, and the measured diffraction amplitudes to determine the projection operator (P_B) in reciprocal space, as described in §2. In this study, for computational convenience the calculations were carried out in space group $P1$, although the true space group of the crystals under study is $P2_12_12_1$. It would be straightforward to adapt the procedure to efficiently account for the presence of crystallographic symmetry and non-orthogonal cell axes, and to construct the real- and reciprocal-space projection operators over the asymmetric region in each space.

Consider first the real-space projection P_A. Generally, as a result of space-group symmetry, there will be a number, denoted n, of molecular envelopes in the unit cell. These envelopes are denoted $U_1, U_2, \ldots, U_n$. In general, some of the molecular envelopes may slightly overlap. The total protein region U is the union of the molecular envelopes, *i.e.* $U = U_1 \cup U_2 \cup \cdots \cup U_n$, and the solvent region, denoted S, is $S = V - U$, where V denotes the region of the unit cell. Let there be N_s grid points in the solvent region S. We define an 'overlap region', denoted O, which is equal to the union of all the intersections of the molecular envelopes, *i.e.* $O = \cup_{\forall(i,j)}(U_i \cap U_j)$. Each grid point within U is associated with a molecular envelope. Consider the molecular symmetry axes which exhibit NCS, *i.e.* those that do not coincide with space-group symmetry axes. Let the resulting NCS be of order R. Then, within each envelope, each grid point j will have $R-1$ equivalent positions, as dictated by the NCS, that will generally not be grid points, which are denoted $j(m)$, where $m = 1, 2, \ldots R$ indexes the NCS operations, and $m = 1$ corresponds to the identity operation. The electron density calculated at position $j(m)$ is denoted x'_{jm}. For $m \neq 1$, the x'_{jm} are calculated by tri-linear interpolation from the electron density at the (maximum of eight) nearest grid points that are in the same envelope as grid point j but not in an overlap region. The number of positions $j(m)$, for fixed j, that are not in the overlap region O, is denoted M_j. With these definitions, and incorporating the solvent flatness constraint, the real-space projection P_A is given by

$$\begin{aligned} P_A x_j &= \frac{1}{M_j} \sum_{\{m:\, j(m) \notin O\}} x'_{jm} \quad \text{for} \quad j \in U - O \\ &= \frac{1}{N_s} \sum_{m \in S} x_m \qquad\qquad \text{for} \quad j \in S. \end{aligned} \tag{7}$$

Consider now the reciprocal-space projection P_B. For reciprocal lattice points where data are measured, this projection involves simply setting the structure-factor amplitudes of the iterate to their measured values and leaving their phases unchanged. At reciprocal lattice points where data are not measured, both the amplitude and the phase of the iterate are left unchanged. Because the iterate is defined in real space, the projection operator also involves a Fourier transform and an inverse Fourier transform operation, and can be written as

$$P_B \mathbf{x} = \mathcal{F}^{-1}[P_{\tilde{B}} \mathcal{F}[\mathbf{x}]], \tag{8}$$

where $\mathcal{F}[\cdot]$ denotes the Fourier transform and the projection $P_{\tilde{B}}$ is defined by

$$P_{\tilde{B}} X_{\mathbf{h}} = \begin{cases} M_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}) & \text{if} \quad \mathbf{h} \in Q, \\ |X_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) & \text{if} \quad \mathbf{h} \notin Q, \end{cases} \tag{9}$$

where $X_{\mathbf{h}}$ denotes the structure factor at reciprocal lattice vector $\mathbf{h}$, *i.e.* $\mathcal{F}[\mathbf{x}] = (X_{\mathbf{h}_1}, X_{\mathbf{h}_2}, \ldots)$, $M_{\mathbf{h}}$ are the measured diffraction-amplitude data after multiplication by the weighting function for the current resolution step and on the same scale as $|X_{\mathbf{h}}|$, $\varphi_{\mathbf{h}}$ denotes the phase of $X_{\mathbf{h}}$, and Q denotes the set of reciprocal lattice points $\mathbf{h}$ where the data are measured (*i.e.* between the minimum and maximum resolutions and excluding any missing data).

3.4. Error metrics

As the iterate $\mathbf{x}_n$ is not an estimate of the electron density as described in §2, it is not appropriate to calculate error metrics based on this quantity. There are various ways of monitoring convergence of these algorithms, and here we use the estimate of the electron density $\hat{\mathbf{x}}$ calculated using (5) at iteration n, and compute conventional crystallographic error metrics as follows. Only the first of these metrics would be applicable to *de novo* phasing applications. However the latter two metrics, which are phase dependent, are useful for benchmarking the performance of the algorithms. Error metrics for iterative projection algorithms are also discussed by Millane & Lo (2013).

The R factor at iteration n is calculated as

$$R_n = \frac{\sum_{\mathbf{h} \in Q} ||\hat{X}_{\mathbf{h}}| - M_{\mathbf{h}}|}{\sum_{\mathbf{h} \in Q} M_{\mathbf{h}}}, \tag{10}$$

where $\hat{X}_{\mathbf{h}}$ is the structure factor of the estimated density $\hat{\mathbf{x}}$ at iteration n. The mean phase error is calculated as

$$\Phi_n = \frac{\sum_{\mathbf{h}} w_{\mathbf{h}} |\hat{\varphi}_{\mathbf{h}} - \varphi^c_{\mathbf{h}}|}{\sum_{\mathbf{h}} w_{\mathbf{h}}}, \tag{11}$$

where $\hat{\varphi}_{\mathbf{h}}$ is the phase of $\hat{X}_{\mathbf{h}}$, $\varphi^c_{\mathbf{h}}$ is the true phase (calculated from the electron density derived from the atomic coordinates), and $w_{\mathbf{h}}$ is the resolution-dependent weighting function (Fig. 1). The similarity between the reconstructed map and the true electron density is measured by the correlation coefficient which is calculated in reciprocal space as (Lunin & Woolfson, 1993)

$$C_n = \frac{\sum_{\mathbf{h} \in Q} |\hat{X}_{\mathbf{h}}| M_{\mathbf{h}} \cos(\hat{\varphi}_{\mathbf{h}} - \varphi^c_{\mathbf{h}})}{\left(\sum_{\mathbf{h} \in Q} |\hat{X}_{\mathbf{h}}|^2 \sum_{\mathbf{h} \in Q} M^2_{\mathbf{h}}\right)^{1/2}}. \tag{12}$$

4. Results

Here we present the results of applying the methods described above to two proteins with ~50% solvent content and fourfold NCS. Both proteins are tetramers with 222 point-group symmetry: tryptophanase from *Proteus vulgaris* (PDB entry 1ax4) and lactate dehydrogenase from *Thermus thermophilus* (PDB entry 2v7p). Inspection of (1) shows that fourfold NCS should be more than sufficient to resolve the phase problem in these cases, and we demonstrate here that the solution can be obtained using iterative projection algorithms.

4.1. *P. vulgaris* tryptophanase

The first molecule used to evaluate algorithm performance is the tryptophanase from *Proteus vulgaris* (Isupov *et al.*, 1998), PDB entry 1ax4. The crystal space group is $P2_12_12_1$ with unit-cell dimensions 115.0 × 118.2 × 153.7 Å. A tetramer with 222 symmetry occupies each asymmetric unit giving fourfold NCS. The solvent content is 50%. Diffraction data are available between 18 and 2.1 Å resolution, with an overall completeness of 97% to 2.1 Å resolution. The data were expanded into space group $P1$ using the crystallographic symmetry. The position and orientation of the NCS axes were derived from the molecular model and are assumed known.

The electron density was sampled on a 110 × 114 × 146 grid, and this grid was used for all resolution steps. The molecular envelope was calculated as described in §3.2, and positioned in the unit cell and replicated by the space-group symmetry. The algorithm was started with electron-density samples within the envelopes chosen independently from a uniform distribution on the interval (0, 1). The actual distribution used was not critical as application of the measured amplitudes in the first iteration sets an approximate scale factor. The difference-map algorithm with $\beta = 0.7$ was applied as described in §§2 and 3. Three different random starting densities were used, but very similar results were obtained for each.

The algorithm was initially run for 200 iterations at each resolution step. Convergence to a good solution was achieved, but inspection of the results indicated that improvements could be made to the algorithm to reduce the number of iterations required. Inspection of the error metrics *versus* iteration showed that for each of the first three resolution steps, the algorithm converged in less than the 200 iterations. For the remaining higher resolution steps, the algorithm converged quickly, in less than 20 iterations, and then started to diverge. The divergent behaviour is a result of the inherent instability of the difference-map algorithm, which is a consequence of its good global search properties as described in §2. In view of this behaviour, two modifications were made to the algorithm which allowed the overall number of iterations to be reduced.

First, as a result of the divergent behaviour, the iterate at the end of a resolution step is not necessarily the best value with which to start the next resolution step. Therefore, at the end of each resolution step, the electron density iterate during that step that has the minimum R factor is used as the starting point for the next resolution step. This gives a better density at the start of each resolution step.

Second, for the higher resolution steps, the algorithm converges quickly and then starts to diverge, so that a large number of iterations is unnecessary. The number of iterations used for each resolution step is therefore determined dynamically by detecting divergence of the algorithm and terminating the iterations for this step at that point. As more iterations are required at the low resolutions where the R factor is larger, the following strategy was used. If the R factor remains greater than 0.4, the full 200 iterations are conducted at that resolution step. If the R factor falls below 0.4, then the iterations at that resolution step are stopped when the behaviour of the R factor indicates that the algorithm is diverging. The assessment of divergence is based on a smoothed version of the R factor as follows. A running average of the R factor is maintained over two preceding contiguous windows of length P iterations each, that are denoted $R_1(n)$ and $R_2(n)$, *i.e.*

$$R_1(n) = \frac{1}{P}\sum_{k=n-P+1}^{n} R_k,$$
$$R_2(n) = \frac{1}{P}\sum_{k=n-2P+1}^{n-P} R_k, \tag{13}$$

where n is the current iteration number. For each resolution step, the first iteration n for which $R_n < 0.4$ and $R_1(n) > R_2(n)$ is taken to be the start of divergent behaviour and the iterations are terminated at that point. The electron density iterate in that step with the minimum R factor is then used to initiate the next resolution step. The value $P = 10$ was found to be effective.

Incorporating these modifications, the algorithm was applied again to the *P. vulgaris* tryptophanase data. Good convergence was again obtained with the total number of iterations being reduced from 2200 to 800. Again, three

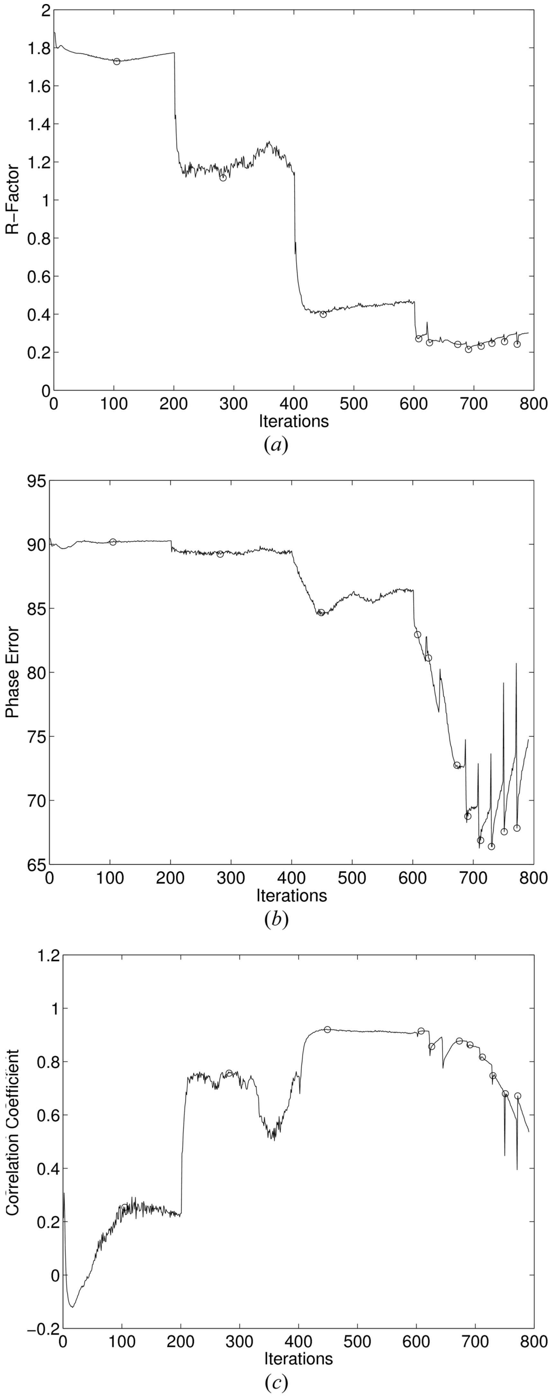


Figure 2
R factor, mean phase error and correlation coefficient (*a*, *b*, *c*) *versus* iteration for *P. vulgaris* tryptophanase. The small circles show the iterations with minimum R factor that are used to initiate the subsequent resolution step.

different starting electron densities were used and very similar results were obtained for each. Plots of the R factor, mean phase error and correlation coefficient are shown in Fig. 2. The iterations with minimum R factor that are used to initiate the next resolution step are indicated by the small circles in the figure.

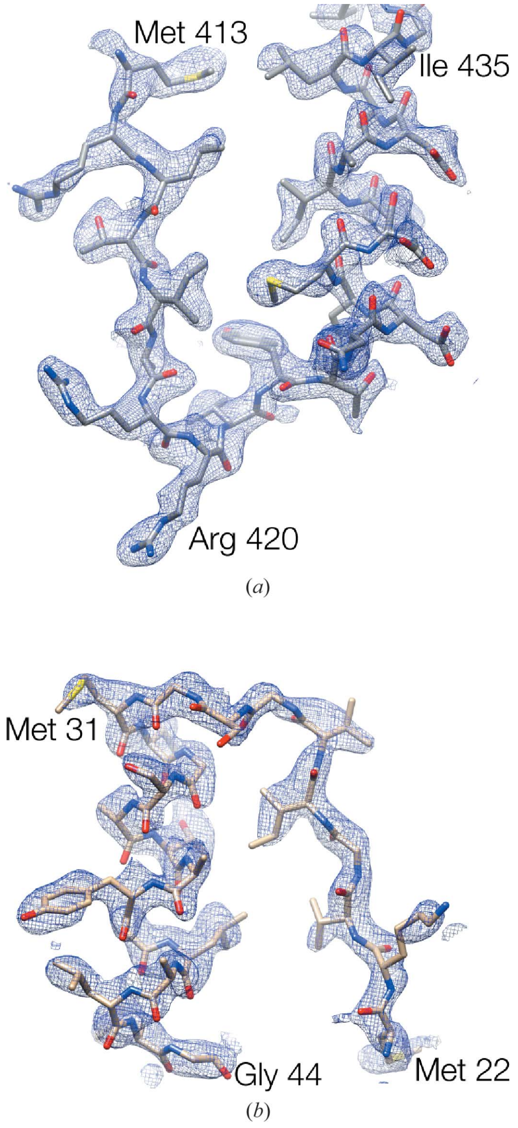
Fig. 2 shows that the error metrics are large for about the first 600 iterations while the algorithm is searching the parameter space for the region of the solution. The R factor is initially large while the correct scale factor is determined, which is initially indeterminate because of the absence of the low-resolution diffraction data. Once the region of the solution is found, the algorithm descends relatively quickly to the solution, at the same time determining the high-resolution phases. Divergent behaviour in the high-resolution steps is evident. The important part of the algorithm is the global search phase during the first ~600 iterations in which the region of the solution is found, despite starting with phases that are far from the correct values. It is this global search ability that sets these kinds of algorithms apart from conventional density-modification algorithms which with poor initial phases would typically stagnate at an early stage and make no progress to the solution.

The final R factor, mean phase error and correlation coefficient are 0.242, 68° and 0.672, respectively, indicating a good solution. Inspection of the reconstructed electron density within the envelope showed that it is clearly interpretable over the majority of the polypeptide chain, and hence suitable for *de novo* model building. The map was used as input to the automated model building procedure ARP/wARP version 7.4 (Langer *et al.*, 2008), which resulted in the successful placement of 83% of the sequence, supporting the above conclusion. An example of the density associated with the C-terminal region of the polypeptide chain is shown in Fig. 3(*a*). The local agreement between the map and model was also analyzed using the program *SFCHECK* (Vaguine *et al.*, 1999). The agreement between the electron density and the model is good almost everywhere, with the regions where agreement is poor largely confined to highly mobile loops on the protein surface. In these regions the atomic displacement parameters of the published model are high, and the electron density is expected to be weak. Overall then, a very satisfactory solution is obtained.

4.2. *T. thermophilus* lactate dehydrogenase

The second molecule studied is another tetrameric protein with 222 point-group symmetry, the lactate dehydrogenase (LDH) from *Thermus thermophilus* (Coquelle *et al.*, 2007), PDB entry 2v7p. This also crystallizes in space group $P2_12_12_1$ with a tetramer in the asymmetric unit. The unit-cell dimensions are 151.5 × 142.9 × 59.6 Å and the solvent content is 50%. Diffraction data are available between 46 and 2.1 Å resolution, with an overall completeness of 98% to 2.1 Å resolution. As in the previous case, the position and orientation of the NCS axes were determined from the molecular model, and are assumed known.

The electron density was sampled on a 146 × 136 × 58 grid and this grid was used at all resolution steps. The molecular envelope was calculated as described in §3.2, and positioned in the unit cell and replicated by the space-group symmetry. The difference-map algorithm was applied starting with a random electron density in the envelopes using the same scheme as for *P. vulgaris* tryptophanase. Three different random starting densities were used and very similar results were obtained for each. The error metrics are shown *versus* iteration in Fig. 4. Similar behaviour is seen as for the case of *P. vulgaris* tryp-



(*a*)

(*b*)

Figure 3
(*a*) Reconstructed electron density associated with amino acids 413–435 of *P. vulgaris* tryptophanase (PDB entry 1ax4), corresponding to the penultimate α-helix of the structure and its preceding β-strand. The deposited atomic model is shown in stick representation, together with an iso-surface of the reconstructed electron-density map. (*b*) Iso-surface of the reconstructed electron density associated with amino acids 22–44 of *T. thermophilus* LDH (PDB entry 2v7p), corresponding to the first α-helix of the structure and its preceding β-strand, together with the deposited atomic model. Figures were prepared using *UCSF Chimera* (Pettersen *et al.*, 2004), with zoning applied to visualize the relevant subregion of the map.

tophanase. The algorithm is in the global search phase for the first ~800 iterations. The mean phase error and the correlation coefficient behave erratically during the first 400 iterations. This behaviour is associated with the global search and subsides after progress is made towards finding the region of the solution. Convergence is obtained at high resolution in 970 iterations and the final values of the error metrics are R = 0.270, Φ = 67° and C = 0.690.

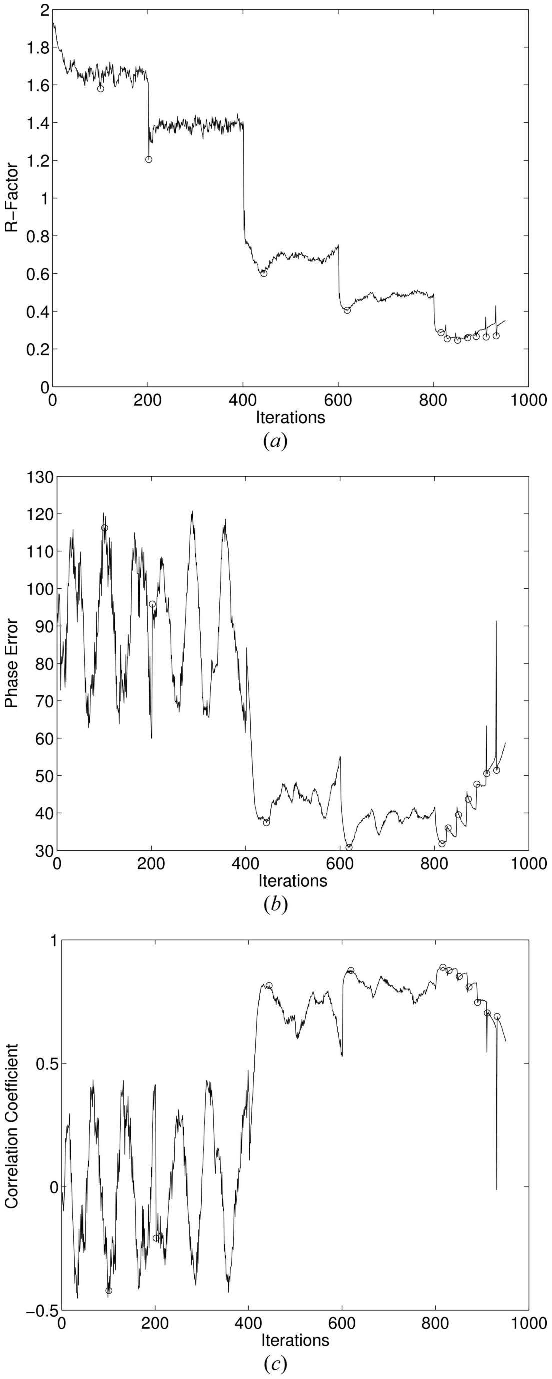


Figure 4
R factor, mean phase error and correlation coefficient *versus* iteration for *T. thermophilus* LDH. The small circles show the iterations with minimum R factor that are used to initiate the subsequent resolution step.

As for *P. vulgaris* tryptophanase, the reconstructed electron-density map was clearly interpretable over the majority of the polypeptide chain. Application of automated model building procedures with ARP/wARP resulted in the successful placement of 92% of the sequence. Electron density associated with the N-terminal region of the polypeptide chain is shown in Fig. 3(*b*). One point of interest in the reconstruction is a poorly defined electron density associated with helix αE (amino acids 112–131). For *T. thermophilus* LDH, this is one of several regions of the structure (amino acids 100–131 and 206–223) which can undergo conformational switching, principally associated with substrate and cofactor binding (Coquelle *et al.*, 2007). Within these regions there are systematic conformational differences between the individual subunits of the LDH tetramer, resulting in departures from the assumed 222 NCS. The breakdown of the NCS degrades the reconstruction despite αE being well ordered in the individual subunits of the tetramer.

5. Discussion

Iterative projection algorithms represent a more sophisticated version of conventional density-modification algorithms that have better global convergence properties. With sufficient, although fairly modest, real-space constraints, they are able to converge to a correct electron density with little or no initial phase information. Application of one of these kinds of algorithm, the difference-map algorithm, to experimental diffraction data from two protein crystals with modest solvent content and fourfold NCS, starting with only a low-resolution molecular envelope and the positions of the NCS axes, leads to high-resolution electron-density maps that are sufficiently accurate for chain tracing. The results confirm the good global convergence properties of these algorithms and their potential for phasing in protein crystallography with minimal additional experimental information. At a minimum, the approach appears competitive with conventional density-modification algorithms. In the presence of low-order NCS, successful high-resolution phasing using these algorithms generally requires initial phase estimates to moderate (5–8 Å) resolution [see, *e.g.*, Nemecek *et al.* (2013)]. We note also that algorithms of the kind we propose can accommodate any real-space constraint, so that inclusion of other features of expected macromolecular electron-density maps would further enhance their effectiveness.

In the results presented, a low-resolution molecular envelope and the position of the NCS axes are needed in order to apply the solvent flatness and NCS constraints at the outset of the procedure. At present, the most straightforward way to generate this information would be to employ conventional experimental phasing techniques such as SAD/MAD and SIRAS. So long as the quality of the phases is sufficient to identify the solvent boundary and locate the NCS axes, iterative projection algorithms can be applied. However, there is potential to incorporate determination of this ancillary information (molecular envelope and position of the NCS axes) directly from the diffraction data into these kinds of algorithms, in which case they would become a viable method for *ab initio* phasing in protein crystallography.

While the current paper was in review, a new paper reporting related work has appeared (He & Su, 2015). These authors further develop the approach of Liu *et al.* (2012), applying the hybrid input–output algorithm to crystals with high solvent content. Their principal new innovation is a scheme for concurrently determining the molecular envelope. This is a significant step, and provides further evidence of the potential of iterative projection algorithms for phasing in protein crystallography.

Acknowledgements

This work was supported by a James Cook Research Fellowship to RPM and a New Zealand Tertiary Education Commission Top Achiever Doctoral Scholarship to VLL.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.
Abrahams, J. P. & Leslie, A. G. (1996). *Acta Cryst.* D**52**, 30–42.
Bracewell, R. N. (1986). *The Fourier Transform and its Applications.* 2nd revised edition. New York: McGraw-Hill.
Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.
Coquelle, N., Fioravanti, E., Weik, M., Vellieux, F. & Madern, D. (2007). *J. Mol. Biol.* **374**, 547–562.
Cowtan, K. (2010). *Acta Cryst.* D**66**, 470–478.
Crowther, R. A. (1969). *Acta Cryst.* B**25**, 2571–2580.
Dodson, E. J. (2001). *Acta Cryst.* D**57**, 1405–1409.
Elser, V. (2003*a*). *J. Opt. Soc. Am. A*, **20**, 40–55.
Elser, V. (2003*b*). *Acta Cryst.* A**59**, 201–209.
Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.
Hao, Q. (2006). *Acta Cryst.* D**62**, 909–914.
Harris, F. J. (1978). *Proc. IEEE*, **66**, 51–83.
He, H. & Su, W. P. (2015). *Acta Cryst.* A**71**, 92–98.
Henderson, R. & Moffat, J. K. (1971). *Acta Cryst.* B**27**, 1414–1420.
Hendrickson, W. A. (2014). *Q. Rev. Biophys.* **47**, 49–93.
Hong, X. & Hao, Q. (2009). *J. Appl. Cryst.* **42**, 259–264.
Isupov, M. N., Antson, A. A., Dodson, E. J., Dodson, G. G., Dementieva, I. S., Zakomirdina, L. N., Wilson, S. K., Dauter, Z., Levedev, A. A. & Harutyunyan, E. H. (1998). *J. Mol. Biol.* **276**, 603–623.
Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nat. Protoc.* **3**, 1171–1179.
Lawrence, M. C. (1991). *Q. Rev. Biophys.* **24**, 399–424.
Liu, Z.-C., Xu, R. & Dong, Y.-H. (2012). *Acta Cryst.* A**68**, 256–265.
Lo, V., Kingston, R. L. & Millane, R. P. (2009). *Acta Cryst.* A**65**, 312–318.
Lo, V. & Millane, R. P. (2010). *Proc. SPIE*, **7800**, 78000N.
Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.
Marchesini, S. (2007). *Rev. Sci. Instrum.* **78**, 011301.
Marks, L. D., Sinkler, W. & Landree, E. (1999). *Acta Cryst.* A**55**, 601–612.
McCoy, A. J. & Read, R. J. (2010). *Acta Cryst.* D**66**, 458–469.
Millane, R. P. (1990). *J. Opt. Soc. Am. A*, **7**, 394–411.
Millane, R. P. & Lo, V. (2013). *Acta Cryst.* A**69**, 517–527.
Millane, R. P. & Stroud, W. J. (1997). *J. Opt. Soc. Am. A*, **14**, 568–579.
Navaza, Q. (2008). *Acta Cryst.* D**64**, 70–75.
Nemecek, D., Plevka, P. & Boura, E. (2013). *Protein J.* **32**, 635–640.
Oszlanyi, G. & Suto, A. (2008). *Acta Cryst.* A**64**, 123–134.
Palatinus, L. (2013). *Acta Cryst.* B**69**, 1–16.
Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.
Plas, J. L. van der & Millane, R. P. (2000). *Proc. SPIE*, **4123**, 249–260.
Rossmann, M. G. (1995). *Curr. Opin. Struct. Biol.* **5**, 650–655.
Sawaya, M. R. (2007). *Methods Mol. Biol.* **364**, 95–120.
Stuart, D. I. & Abrescia, N. G. A. (2013). *Acta Cryst.* D**69**, 2257–2265.
Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22–37.
Thumiger, A. & Zanotti, G. (2009). *Croat. Chem. Acta*, **82**, 421–432.
Tong, L. & Rossmann, M. G. (1997). *Methods Enzymol.* **276**, 594–611.
Trapani, S., Schoehn, G., Navaza, J. & Abergel, C. (2010). *Acta Cryst.* D**66**, 514–521.
Urzhumtsev, A. G. & Podjarny, A. W. (1995). *Acta Cryst.* D**51**, 888–895.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
Waser, J. & Shomaker, V. (1953). *Rev. Mod. Phys.* **25**, 671–690.
Weichenberger, C. X. & Rupp, B. (2014). *Acta Cryst.* D**70**, 1579–1588.
Xiong, Y. (2008). *Acta Cryst.* D**64**, 76–82.