Version

This is the Accepted Manuscript version. This version is defined in the NISO recommended practice RP-8-2008 http://www.niso.org/publications/rp/

Suggested Reference


Copyright

Items in ResearchSpace are protected by copyright, with all rights reserved, unless otherwise indicated. Previously published items are made available in accordance with the copyright policy of the publisher.

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

For more information, see General copyright, Publisher copyright, Sherpa Romeo.
Iterative projection algorithms for \textit{ab initio} phasing in virus crystallography

Victor L. Lo$^a$, Richard L. Kingston$^b$, Rick P. Millane$^{a,*}$

$^a$Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand
$^b$School of Biological Sciences, The University of Auckland, Auckland, New Zealand

Abstract

Iterative projection algorithms are proposed as a tool for \textit{ab initio} phasing in virus crystallography. The good global convergence properties of these algorithms, coupled with the spherical shape and high structural redundancy of icosahedral viruses, allows high resolution phases to be determined with no initial phase information. This approach is demonstrated by determining the electron density of a virus crystal with 5-fold non-crystallographic symmetry, starting with only a spherical shell envelope. The electron density obtained is sufficiently accurate for model building. The results indicate that iterative projection algorithms should be routinely applicable in virus crystallography, without the need for ancillary phase information.

\textbf{Keywords:} virus, crystallography, phasing, \textit{ab initio}, iterative projection algorithm

1. Introduction

Determination of macromolecular structures by x-ray crystallography is plagued by the phase problem: the phases of the diffracted x-rays cannot be measured and must be inferred using ancillary techniques; the phases being required to compute the electron density. Generally, the phase problem is solved experimentally using the methods of isomorphous replacement or anomalous dispersion,
or computationally using the method of molecular replacement. Each of these methods can present difficulties, however, and the resulting phases may not be accurate enough to produce an interpretable electron density map. The method of molecular replacement requires knowledge of a related structure, and therefore is not useful for the determination of novel structures.

In x-ray crystallography of icosahedral viruses the difficulties presented by the phase problem are alleviated to some degree by the high internal symmetry of the viral particle. If elements of the particle symmetry do not coincide with the crystal space group symmetry, i.e. are “noncrystallographic,” then constraints are automatically placed on valid phase sets [3, 1, 15]. Because of the crystallographic restriction, crystals of icosahedral viruses always exhibit at least 5-fold noncrystallographic symmetry (NCS). Depending on the setting of the particle in the unit cell, NCS of 60-fold or even higher order can be observed.

NCS has been used to advantage in virus crystallography, through NCS averaging and phase extension. Cryo-EM maps of either the intact virus or of viral capsomeres may be used to obtain a set of initial, low resolution phase estimates. Alternatively, initial phases may come from the x-ray structure of a homolog, or from other sources. The initial phases are refined by an iterative density modification protocol incorporating NCS and solvent flatness as constraints [32]. Additional diffraction amplitudes in a thin resolution shell are then incorporated and the phases refined again by the same procedure. The resolution is extended step-wise in this fashion, out to the maximum resolution of the diffraction data. In favorable cases, the resulting phases at high resolution are accurate enough to calculate an electron density map that is suitable for model building.

Early successful applications of this approach included the following. Determination of the structure of bacteriophage MS2 used 10-fold NCS averaging, initial molecular replacement phases at 13Å resolution, and phases from two heavy-atom derivatives at 8.8Å resolution [27]. Structure determination of canine parvovirus (CPV) used 60-fold NCS averaging and initial single isomorphous replacement phases at 8Å resolution [26]. The structure of Nudaurelia
capensis ω virus was determined using 60-fold NCS averaging and data from one isomorphous heavy-atom derivative as well as partial model building [20].

More recently, cryo-EM has become a significant source of initial phase information. The structure of the human Hepatitis B virus capsid was determined using 30-fold NCS averaging, starting with cryo-EM derived phases at 8Å resolution [31]. The structure of human adenovirus 2 penton was determined using 60-fold NCS averaging and molecular replacement phases from a related 15Å resolution cryo-EM model [33]. The structure determination of a picobirnavirus used 30-fold NCS averaging, starting with phases from a cryo-EM map at 20Å resolution [4]. In all of these cases, some form of experimental phase information was used to obtain a solution.

The potential of ab initio phase determination in icosahedral virus crystallography has long been recognised [22], though seldom realised in practice. In the case of CPV, retrospective analysis showed that if the position of the particle had been accurately known, then, starting with an initial spherical envelope, phases of similar accuracy to the to SIR phases at 9Å resolution could have been obtained [26]. This presumably would have allowed refinement to high resolution, without the use of the SIR phases. This demonstrates the feasibility of ab initio phasing with 60-fold NCS. Miller et al. (2001) [19] used an initial spherical envelope and genetic algorithms to determine the phases of a form of poliovirus to 20.5Å resolution, and the phases were subsequently extended to high resolution using conventional symmetry averaging and phase extension. However, it was necessary to incorporate very low resolution (300Å) diffraction data and the order of the NCS was high (30-fold). This approach has not developed into a general technique. Taka et al. (2005) [24] conducted simulations of the reconstruction of icosahedral viruses starting with a spherical shell model and using conventional density modification. Although some success was achieved, the resolution was limited (maximum resolution between 4.3 Å and 10Å), and some cases required incorporation of some calculated, rather than measured, diffraction amplitudes. This approach, also, has not developed into a general technique.
In summary then, despite the success of conventional symmetry averaging
and phase extension methods in virus crystallography, it has generally been nec-
essary, even with high order NCS, to start the phase determination procedure
with an experimentally obtained low to medium resolution phase set. Yet it
is well established that the structural redundancy present places considerable
constraints on the phases. We have shown recently [17, 16] that quite low-order
structural redundancy is theoretically sufficient to render the solution to the
phase problem unique, i.e. to constrain the phases to their correct values in
the absence of any other experimental information. For example, for a virus
crystal with 5-fold NCS and 40% solvent, the phase problem is overdetermined
by a factor of 4. Therefore, it should be possible to determine high resolu-
tion phases in virus crystallography with information on only the nature of the
NCS, and what is needed is a phasing algorithm that has a sufficiently large
radius of convergence to find the solution [17]. Algorithms that have a larger
radius of convergence than current electron density modification algorithms are
therefore of interest. We have described elsewhere “iterative projection algo-
rithms,” a class of electron density modification-like algorithms that have good
global convergence properties, and their potential for ab initio phasing in pro-
tein crystallography [17]. These kinds of algorithms are particularly suitable for
ab initio phasing in virus crystallography since (1) reasonably high-order NCS
is always present, (2) the positions of the NCS axes are often easily determined,
and (3) a spherical shell usefully approximates the molecular envelope at low
resolution. These characteristics allow a potentially straightforward and effective application of iterative projection algorithms in virus crystallography. Here
we demonstrate the power of this approach by reconstructing the electron den-
sity of a virus crystal with 5-fold NCS from the experimental diffraction data
and only an initial spherical shell envelope. The resulting electron density is
sufficiently accurate for model-building.

This paper is organized as follows. In Section 2 we briefly review iterative
projection algorithms. Some specific details of the implementation of the phas-
ing algorithm are described in Section 3. Results of the application to ab initio
phasing of an icosahedral virus are described in Section 4. Concluding remarks are made in Section 5.

2. Iterative projection algorithms

Conventional phase determination in virus crystallography proceeds by phase extension to increasingly high resolution. Within each phase extension step, the phase set is refined by iterating between real space and reciprocal space, applying NCS and solvent flatness constraints in the former, and forcing agreement with the measured structure factor amplitudes in the latter. It is convenient to represent the electron density at the grid points in the unit cell as the elements of a vector denoted $\mathbf{x}$. The $j$th element of $\mathbf{x}$, $x_j$, is then the electron density at the $j$th grid point in the unit cell. The operations in real and reciprocal space described above are denoted by the operators $P_A$ and $P_B$, respectively. These are termed “projections,” that operate on the vector $\mathbf{x}$. Therefore, applying NCS averaging and/or solvent flattening to an electron density $\mathbf{x}$, gives the new density $\mathbf{y}$, where $\mathbf{y} = P_A \mathbf{x}$. Similarly, taking the Fourier transform of the density $\mathbf{x}$, replacing the structure factor amplitudes with their measured values (but keeping the phases unchanged), and taking the inverse Fourier transform, gives the new density $\mathbf{z}$, where $\mathbf{z} = P_B \mathbf{x}$. In this formalism then, one iteration of conventional iterative phasing, or density modification, can be written succinctly as the “update rule”

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

where $\mathbf{x}_n$ denotes the electron density at iteration $n$. This algorithm is sometimes referred to as the error-reduction algorithm in the image processing literature [6]. Since the algorithm is recursive, it must be initialised with an initial electron density or an initial phase set.

However, although the algorithm Eq. (1) is easy to understand and simple to apply, it has poor global convergence properties, i.e. convergence to the correct solution is achieved only if the initial phase set is close enough to the correct
phase set [5, 17]. We expect, therefore, that this poor global convergence is a
contributing factor to the difficulties of ab initio phasing, even with reasonably
high order NCS. With a sufficiently poor initial phase set, the algorithm will
generally converge to a phase set, or electron density, that is unrelated to the
correct values. This is sometimes referred to as “stagnation.” The conventional
density modification algorithm Eq. (1) is therefore not well-suited to ab initio
phasing, where one is starting with no initial phase information.

There are, however, iterative algorithms that have good global convergence
properties. If a unique solution exists, these algorithms often locate that so-
lution, independently of where they are initialised [5, 13, 17]. Such algorithms
therefore have potential for ab initio phasing in virus crystallography, starting
with little, or no, initial phase information. These algorithms are generally
called “iterative projection algorithms” [5, 17]. Iterative projection algorithms
are recursive, and they make use of the same projection operators $P_A$ and $P_B$
as in Eq. (1), but the operators are combined in different ways than in Eq. (1)
to form the update rule. There are a variety of such algorithms in use and the
reader is referred to various reviews [13, 17].

In this work we use the so-called “difference map algorithm” [5], which is
a flexible algorithm that has good global convergence properties. The update
rule for the difference map algorithm is

$$x_{n+1} = x_n + \beta \left[ P_A F_B(1/\beta)x_n - P_B F_A(-1/\beta)x_n \right].$$

(2)

The operators $F_A(-1/\beta)$ and $F_B(1/\beta)$ are referred to as relaxed projection operators onto the constraint sets $A$ and $B$, respectively, and are defined in terms of the projections operators $P_A$ and $P_B$ by

$$F_A(-1/\beta)x = P_Ax + (-1/\beta)(P_Ax - x)$$
$$F_B(1/\beta)x = P_Bx + (1/\beta)(P_Bx - x).$$

(3)

The difference map algorithm has the single parameter $-1 < \beta < 1$, which
is usually fixed, and values $\beta \approx 0.7$ are often suitable. The relaxed projection
operator in Eq. (3) under- or over-projects, depending on the value of $\beta$, and, by
itself, is reminiscent of, for example, \(\gamma\)-correction \cite{17}. The update rule Eq. (2) uses only the basic projection operators \(P_A\) and \(P_B\) as in conventional density modification, and is only slightly more complicated algorithmically than Eq. (1). The computational cost is only about twice that of the conventional algorithm per iteration. The algorithm can therefore be applied rather straightforwardly in virus crystallography. The theoretical rationale for the difference map algorithm is described in Ref. \cite{5}.

We refer to \(x_n\), in both Eq. (1) and Eq. (2), as the “iterate.” For iterative projection algorithms, the iterate is an auxiliary function that is used to search the parameter space. For the error reduction algorithm, Eq. (1), the iterate is an estimate of the electron density. For the difference map algorithm, Eq. (2), the iterate is not an estimate of the electron density, but when the algorithm has converged, i.e. \(x_{n+1} \approx x_n\), the solution, denoted \(\hat{x}\), can be calculated as

\[
\hat{x} = P_A F_B (1/\beta) x_n^*,
\]

where \(x_n^*\) denotes the value of \(x_n\) at convergence \cite{5,17}. Alternatively, after the difference map algorithm has converged, a few cycles of the conventional update rule Eq. (1) will quickly converge to the solution.

A characteristic of iterative projection algorithms is that they can be inherently unstable. This is a consequence of their ability to escape from local minima. They can, for example, drift away from a solution, and measures may need to be taken to arrest this behaviour. This is discussed further in Section 3.1.

There has been some exploratory application of iterative projection algorithms in protein and virus crystallography. Millane and Stroud (1997) \cite{18} and van der Plas and Millane (2000) \cite{28} adapted the hybrid input-output (HIO) algorithm \cite{6} to incorporate a NCS constraint and applied it to reconstruction of an icosahedral virus at medium resolution using simulated data. Lo et al. (2009) \cite{9} applied the difference map algorithm to determination of molecular envelopes in protein crystals from simulated solvent contrast variation data. Lo and Millane (2010) \cite{11} applied the difference map algorithm to reconstruction of a
previously solved icosahedral virus using experimental data, which gave reasonably good electron density maps. Liu et al. (2012) [8] and He and Su (2015) [7] have applied the HIO algorithm with a solvent flatness constraint to a number of solved proteins with high (> 65%) solvent content. Recently, Lo et al. (2015) [10] applied the difference map algorithm to two solved protein structures with lower (~50%) solvent content by incorporating a NCS constraint. Here we describe in detail application of the difference map algorithm to ab initio phasing of a previously solved icosahedral virus using the experimental diffraction data, and starting with only a spherical shell envelope. We show that the resulting electron density map is sufficiently accurate for model building.

3. Methods

The basic algorithm we use here is the difference map algorithm as described in Section 2. However, two important additional components of our phasing procedure are weighting of the diffraction data during phase extension, and refinement of the molecular envelope. These and other aspects of the algorithm are described in this section.

3.1. Resolution extension

Phase, or resolution, extension by classical density modification in virus crystallography usually involves incorporating the higher resolution diffraction amplitudes in a thin shell, often only one reciprocal lattice spacing in width, at each phase extension step. This is equivalent to applying a discontinuous spherical weighting function to the data, that is uniform within the sphere and zero outside it. We have found it advantageous to employ instead a smooth weighting function that decreases monotonically with resolution. At each resolution extension step, the weighting function is changed so that the weight applied to the higher resolution data is increased. This revised weighting scheme has two effects. The smooth, rather than sharp, resolution cutoff reduces ripples in the electron density and has the effect of reducing the number of iterations...
required in each resolution extension step. Furthermore, the number of resolution extension steps needed is reduced significantly from what it would be if the extensions were made one reciprocal lattice spacing at a time. The overall effect is a substantial reduction in the total number of iterations required.

The weighting function that we use has a (one-sided) Gaussian dependence on the reciprocal of resolution, and the diffraction amplitude data are multiplied by this weight function. At the beginning at each resolution extension step, the width of the Gaussian function is increased, thus giving a larger weight to the higher resolution data. In the experiments described here, 11 resolution extension steps were used with the Gaussian weighting function having half-heights at 30, 20, 12, 8, 6, 5, 4, 3, 2.5, and 2 Å resolution. In the final resolution step, the weighting function is constant, i.e. the diffraction data are unweighted.

Two other modifications were also made to the usual phase extension protocol, which allowed the overall number of iterations required to be further reduced. The first modification ameliorates the effect of possible instability as mentioned in Section 2. Rather than initiating the electron density at the beginning of a phase extension step with the density at the end of the previous step, it is set to the electron density iterate in the previous step that has the minimum R-factor. The result is a better density at the start of each phase extension step, hastening convergence. Second, the number of iterations used in each phase extension step is determined dynamically, based on the behaviour of the algorithm in that step, as follows. If the R-factor remains greater than 0.4 (generally the case for the low resolution steps), 200 iterations are conducted in that phase extension step. If the R-factor falls below 0.4 in a phase extension step, then the iterations in that step are terminated when the algorithm is assessed to be no longer converging. Convergence is evaluated by calculating a running average of the R-factor over the 10 previous iterations and comparing it to the running average over the 10 iterations prior to that. If the average R-factor has increased, the algorithm is deemed to be no longer converging and the iterations are terminated at that point in that phase extension step.
3.2. Envelope definition and refinement

We assume that no initial phase estimates are available to start the phasing process. The algorithm is therefore effectively started with random phases. However, in order to apply the NCS and solvent flatness constraints at the beginning of the process, an initial envelope is required within which to apply the NCS, and outside of which to enforce solvent flatness. In the absence of any phase information, the overall particle dimensions could be estimated using dynamic light scattering (DLS), small angle X-ray solution scattering (SAXS) or transmission electron microscopy (TEM). Chapman et al. [2] describe methods for estimating the overall virus dimensions from the diffraction data. We assume here only that initial estimates of the capsid inner and outer radii are available.

In many cases, the particle orientation within the crystal can be deduced from inspection of the self-rotation function [25, 23]. Often, including the example described in the next section, both particle orientation and position are fixed by the space group. For the case of a particle in a general position in the unit cell, fixing the particle location with sufficient accuracy for NCS averaging, in the absence of phase information, can be difficult.

The spherical shell envelope is a good approximation only at low resolution, and significant advantage is obtained if the envelope is refined as the phase refinement proceeds and the resolution is increased [30]. There are various ways of doing this [30, 14], and here we used the procedure described below.

The envelope is updated (and departs from the initial spherical shell) at each resolution step, based on the current estimate of the electron density as follows. The updated envelope is used in the NCS averaging and the solvent flattening projection at each iteration. Three regions are defined in the unit cell. One of these is the region exterior to the capsid envelopes and is denoted \( V_e \). The second is the interior region of the capsid envelopes which is denoted \( V_i \). The third region is that of the capsids themselves, which is denoted \( V_c \).

At the beginning of each resolution extension step, the envelopes are updated by updating the region \( V_c \) (and hence also \( V_e \) and \( V_i \)), based on the current estimate of the electron density. The region \( V_c \) contains solvent, \( V_i \) generally
contains solvent and disordered nucleic acid, and $V_c$ contains protein. The electron densities in regions $V_e$ and $V_i$ are assumed to be flat, and are denoted $\rho_e$ and $\rho_i$, respectively. Since the value of $F_{000}$ is unknown, there is one unknown degree of freedom that corresponds to the average electron density in the unit cell. We fix this degree of freedom by setting $\rho_e$ to a fixed, constant value that is not changed during the phase determination procedure. The value of $\rho_i$ is updated at each iteration as described in the next section. Initially, the regions $V_e$, $V_i$ and $V_c$ are defined by the initial spherical shell envelope. This initial envelope is conservative in the sense that $V_e$ contains only solvent, $V_i$ contains only solvent/nucleic acid, and $V_c$ will initially contain protein, solvent and nucleic acid. The objective of the envelope refinement is to reduce the size of $V_c$ as the resolution increases so that it tends to contain only protein. For the envelope refinement, a density $x'$ is defined, whose components $x'_j$ are given by

$$x'_j = \begin{cases} 0 & \text{for \ } j \in V_e \ \text{or \ } j \in V_i \\ |x_j - \rho_e| & \text{for \ } j \in V_c \ \text{and \ } R_j > \bar{R} \\ |x_j - \rho_i| & \text{for \ } j \in V_c \ \text{and \ } R_j < \bar{R}, \end{cases}$$

(5)

where $R_j$ is the distance of grid point $j$ from the center of the particle, and $\bar{R}$ is the average of the inner and outer radii of the initial spherical shell envelope. The density $x'$ is then smoothed by multiplying the corresponding structure factors by a Gaussian window with half-height at 10Å resolution, giving a density denoted $x''$. The updated region $V_c$ is then defined as the set of grid points that contain the $\alpha f N$ largest values of $x''$, where $\alpha$ is a constant slightly larger than unity, $f$ is the protein content of the crystal, and $N$ is the number of grid point in the unit cell. The regions $V_e$ and $V_i$ are updated accordingly. This procedure was found to be effective and we used $\alpha = 1.1$. This then defines the updated envelope, which evolves away from the initial spherical shell as the resolution increases.
3.3. Projections

Solvent flatness and NCS constraints determine the projection operator \( P_A \) in real space, and the measured diffraction amplitudes determine the projection operator \( P_B \) in reciprocal space. For computational convenience, the calculations were carried out here in space group P1. It would be straightforward to adapt the procedure to utilize the crystallographic symmetry, and to perform the calculations over the asymmetric regions in real and reciprocal space.

The real space projection is performed in the usual way, and the electron density at grid point \( j, x_j \), is updated as

\[
P_A x_j = \begin{cases} 
\rho_e & \text{for } j \in V_e \\
\bar{\rho}_i & \text{for } j \in V_i \\
\bar{x}_j & \text{for } j \in V_c
\end{cases}
\]

where \( V_e, V_i \) and \( V_c \) are the current envelopes. In (6), \( \bar{\rho}_i \) is equal to the average of the current density in region \( V_i \), and \( \bar{x}_j \) is the current density averaged over the grid points that are NCS-related to grid point \( j \) (calculated by trilinear interpolation). Equation (6) then enforces the NCS and solvent flatness constraints using the current envelope.

The reciprocal space projection \( P_B \) is also performed in the usual way. Since the iterate \( x_n \) is defined in real space, the projection operator \( P_B \) involves Fourier transforming the iterate, updating the structure factor amplitudes, and inverse Fourier transforming to obtain the projected iterate. For reciprocal lattice points where diffraction data are measured, the structure factor amplitudes are replaced by their (weighted) measured values (and the phases are unchanged). At reciprocal lattice points where data are not measured, both the amplitude and phase of the structure factors are unchanged from their current values. The projection operators \( P_A \) and \( P_B \) are then incorporated into one iteration of the difference map algorithm using Eqs. (2) and (3).
4. Results

Here we present the results of applying the methods described above to \textit{ab initio} phasing of the melon necrotic spot virus (MNSV) \cite{29}, PDB code 2zah. The virus has icosahedral symmetry and crystallized in space group \textit{I}2\textit{3}, with two particles in the unit cell. The unit cell dimensions are 375.0 \times 375.0 \times 375.0\text{Å}. With this packing arrangement there is exact 5-fold NCS present, associated with one of the rotational symmetry axes of the icosahedral particle. Experimental diffraction data are available between 267 and 2.8\text{Å} resolution, with an overall completeness of 99\%. All data beyond a lower resolution limit of 150\text{Å} were used here. The data were expanded into space group \textit{P}1 using the crystallographic symmetry. The space group and icosahedral symmetries together fix the position and orientation of the 5-fold rotational NCS axes.

The electron density was sampled on a 268 \times 268 \times 268 grid, and this grid was used for all resolution extension steps. The initial spherical shell molecular envelope had with inner and outer diameters of 210\text{Å} and 340\text{Å}, respectively, which were taken from the known virion dimensions. This initial envelope was 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 immediately re-scales the density. The difference map algorithm with $\beta = 0.7$ was applied as described in Sections 2 and 3.

Some care is needed in calculating error metrics in iterative transform algorithms as the iterate $x_n$ is not necessarily an estimate of the electron density, as described in Section 2, and it is not therefore appropriate to use this quantity to calculate error metrics \cite{17}. Conventional metrics can be calculated however if the electron density estimate $\hat{x}$, calculated using Eq. (4) at iteration $n$, and its Fourier transform, are used. We use the R-factor ($R$), mean phase error ($\Phi$), and map correlation coefficient ($C$) \cite{12} to monitor convergence \cite{17, 10}.

Applying the difference map algorithm using the protocol described above,
convergence to a good solution was achieved in 870 iterations. Three 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 versus iteration are shown in Fig. 1. The resolution extension steps, i.e. the iterations where the width of the Gaussian weighting function is increased, are shown by the vertical lines in the figure. The iterations with minimum R-factor that are used to initiate the next phase extension step are indicated by the small circles in the figure.

Inspection of Fig. 1 shows that all error metrics remain large until the end of phase extension step 3 (weighting function half-height at 12\AA{} resolution), at iteration 470. At this point, the R-factor is 0.45, the mean phase error 54° and the correlation coefficient 0.54. These values indicate that the solution is in a local minimum, not particularly close to the correct solution. During the next phase extension step (weighting function half height at 8\AA{} resolution), the error metrics increase markedly for some time and then improve rather dramatically to $R = 0.26$, $\Phi = 43^\circ$ and $C = 0.72$ at iteration 670. The iterate has escaped from the local minimum, passing through a region of the parameter space corresponding to a poor solution, and is now closer to the true solution. During the next phase extension step, the iterate drops into a region closer to the solution. As the high resolution data is introduced in subsequent phase extension steps, the iterate remains in the vicinity of the solution and continues to improve. The final error metrics against the full unweighted dataset, at iteration 870, are $R = 0.19$, $\Phi = 32^\circ$ and $C = 0.83$, indicating that a good solution has been located.

The observed behaviour highlights the key elements of iterative projection algorithms. During the first 670 iterations, the algorithm is in a global search phase during which it explores a large region of the parameter space, starting from a random initial position, entering and exiting from local minima, passing through regions of poor agreement, and finishing at a density in the vicinity of the correct density. The iterate then appears to move in a smaller region of the parameter space and moves more smoothly to the solution as the resolution is
increased. It is this global search capability of iterative projection algorithms that sets them apart from conventional algorithms, that will typically stagnate near a local minimum when started with random phases.

Consistent with the global error metrics, inspection of the reconstructed electron density shows that the envelope definition has worked correctly (Fig. 2(A)). The map is clearly interpretable over the majority of the polypeptide chain (e.g. Fig. 2(B)), and is hence suitable for de novo model building. Overall then, a very satisfactory solution is obtained.

If the error reduction algorithm is used from the outset, no progress is made to the solution, as expected. It is not clear from the error plots in Fig. 1 when the iterate enters a local convergence region. Once this has occurred, a local search algorithm, such as the error reduction algorithm, should be capable of locating the solution. The error reduction algorithm was run starting from various iterations after running the difference map algorithm. It was found, however, that convergence is obtained only if the switch to the error reduction algorithm occurred in the final resolution step. This indicates that the presence of local minima is significant throughout almost all of the procedure, and that a global search algorithm is needed out to high resolution.

An initial definition of the molecular envelope is required in order to start the phase determination procedure. Although an accurate initial envelope is not needed, it is preferable for its dimensions to slightly exceed the true dimensions to avoid flattening density in the protein region at the beginning of the procedure. Since determination of the dimensions of the initial envelope will be prone error, we also ran our algorithm starting with larger initial spherical shell envelopes, i.e. increasing the outer diameter and decreasing the inner diameter. The performance of the algorithm was unaffected with the outer diameter increased by up to 15Å (to 355Å) and the inner diameter decreased by up to 10Å (to 200Å). For example, with initial inner and outer diameters of 200Å and 355Å, respectively, the final error metrics \( R = 0.20, \Phi = 35^\circ \) and \( C = 0.81 \), obtained in 650 iterations, are almost identical to those obtained with the more accurate initial envelope. This indicates that a highly accurate initial envelope
Figure 1: R-Factor, mean phase error and correlation coefficient (a, b, c) versus iteration for MNSV. The vertical lines show the resolution extension steps, i.e. the iterations at which the width of the Gaussian weighting function is increased. The small circles show the iterations with minimum R-factor that are used to initiate the subsequent phase extension step.
Figure 2: (A) Electron density isosurface for the entire unit cell (space group I23). The top right octant of density has been erased to make the interior of the centrally-located icosahedral particle visible. A magenta-cyan color gradient extends from the front to the rear of the unit cell. (B) Electron density iso-surface associated with the P domain of the viral capsid protein. The iso-surface is displayed as a mesh, with the coordinates of the published atomic model (PDB ID: 2ZAH) displayed in stick representation. Zoning was applied to visualize the relevant subregion of the map. Figures were generated using UCSF Chimera [21].
is not needed for successful \textit{ab initio} phasing of diffraction data from icosahedral virus crystals using iterative projection algorithms.

5. Summary

The high degree of structural redundancy in icosahedral viruses has long suggested that \textit{ab initio} phasing in virus crystallography should be possible. As a result of the redundancy, there is generally enough information in the diffraction amplitudes to uniquely determine the phases without any additional information. Despite this, current electron density modification algorithms have so far not been effective in determining high resolution phases without the availability of some initial, low to medium resolution phase information. Iterative projection algorithms are a more sophisticated version of conventional density modification algorithms that have better global convergence properties, and offer the possibility of \textit{ab initio} phasing in virus crystallography. Application of the difference map algorithm to experimental diffraction data from a virus crystal with 5-fold NCS, starting with only a spherical shell molecular envelope, results in a high resolution electron density map that is sufficiently accurate for chain tracing. The results confirm the good global convergence properties of these algorithms when started with random phases, and their potential for \textit{ab initio} phasing in virus crystallography.

6. Acknowledgements

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

References


