Acta Cryst A FOUNDATIONS ADVANCES

ISSN 2053-2733

Uniqueness of the macromolecular crystallographic phase problem

Rick P. Millane* and Romain D. Arnal

Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand. *Correspondence e-mail: rick.millane@canterbury.ac.nz

Received 8 June 2015
Accepted 17 August 2015

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

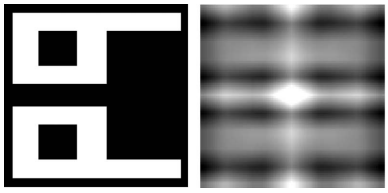
Keywords: uniqueness; phase problem; *ab initio* phasing; macromolecular crystallography.

Uniqueness of the phase problem in macromolecular crystallography, and its relationship to the case of single particle imaging, is considered. The crystallographic problem is characterized by a constraint ratio that depends only on the size and symmetry of the molecule and the unit cell. The results are used to evaluate the effect of various real-space constraints. The case of an unknown molecular envelope is considered in detail. The results indicate the quite wide circumstances under which *ab initio* phasing should be possible.

1. Introduction

The phase problem is of key importance in macromolecular crystallography, and phase determination can be a limiting step in protein structure determination. In the absence of the applicability of direct methods (due to the large number of atoms and the moderate resolution of the diffraction data), information in addition to the structure-factor amplitudes is required to retrieve the phases (Drenth, 1994). For example, additional experimental diffraction data, such as from modified crystals or from anomalous dispersion, or knowledge of a known related structure provide phase information. Alternatively, or as well as, *a priori* real-space information [such as generic properties of protein electron densities, or non-crystallographic symmetry (NCS)] also constrains the phases. In this article we consider the case of *ab initio* phasing using real-space information, in the absence of any experimental phase information. A question of theoretical and practical importance in this case is: to what extent does real-space information constrain the phases and, importantly, what real-space information is sufficient to render the solution to the phase problem unique? *Ab initio* phasing in protein crystallography has so far met with limited success, current methods being based on direct methods utilizing random placement of atoms or secondary structure fragments, and which are generally applicable only to small proteins with reasonably high resolution diffraction data (Sheldrick *et al.*, 2011; Millán *et al.*, 2015). However, approaches based on iterative projection algorithms (Elser, 2003*a*; Marchesini, 2007; Millane & Lo, 2013) have recently shown considerable promise (Liu *et al.*, 2012; He & Su, 2015; Lo *et al.*, 2015), and these results prompt a more definitive analysis of the uniqueness question in macromolecular crystallography.

Crowther (1969) and Bricogne (1974) considered the effect of structural redundancy on constraining the phases in terms of the number of observations and the number of parameters describing the electron density of the subunit from which the electron density in the unit cell is built. Millane (1993) gave a quantitative assessment of uniqueness for macromolecular



crystallography as a function of the shape of the molecular envelope and the order of any NCS present. The results of Miao *et al.* (1998) indicated that, for a 'full' unit cell, the structure amplitudes alone underdetermine the phase problem by a factor of a half.

The phase problem for a single, isolated object has been considered separately in the literature. It has been shown that the phase problem for a single object in two or more dimensions has a unique solution (Fienup, 1978; Bruck & Sodin, 1979; Bates, 1984; Barakat & Newsam, 1984; Millane, 1990). This was considered further by Millane (1996) who showed that the problem is better determined as the number of dimensions is increased. Elser & Millane (2008) characterized the nature of the problem by defining a 'constraint ratio', and showed that this can be expressed as a function of only the object shape.

In this paper we make connections between uniqueness properties of the phase problem for single objects and for a crystal. We obtain an expression for the constraint ratio for crystals that allows the effect of different real-space constraints to be evaluated. This expression is applied to the cases of a restricted molecular envelope, and crystallographic and non-crystallographic symmetry. The case of an unknown molecular envelope is considered in detail. The results assist in understanding the nature of the macromolecular crystallographic phase problem and the potential for *ab initio* phasing.

2. Uniqueness for a single object

Consider first the phase problem for a single, isolated object, as in, for example, single particle imaging or astronomy *etc.* In this case, the Fourier amplitude is measured continuously in Fourier, or reciprocal, space, *i.e.* there is no Bragg sampling. It is well known that in this case, as long as the object is in two or more dimensions, the solution to the phase problem is unique (Fienup, 1978; Bruck & Sodin, 1979; Bates, 1984; Barakat & Newsam, 1984). Uniqueness can be characterized by the constraint ratio, denoted Ω, which is equal to the number of independent data contained in the amplitude data divided by the number of parameters describing the object (at the resolution of the data) (Elser & Millane, 2008). Note that the number of data cannot be increased indefinitely by finer sampling of the amplitude, since the number of *independent* data is limited by the sampling theorem. The constraint ratio can be expressed in the form (Elser & Millane, 2008)

$$\Omega = \frac{|\mathcal{A}|}{2|\mathcal{S}|}, \tag{1}$$

where $\mathcal{S}$ denotes the support of the object (*i.e.* the region occupied by the object), $\mathcal{A}$ denotes the support of the autocorrelation of the object (or of $\mathcal{S}$), and $|\cdot|$ denotes the volume. The phase problem is well determined if $\Omega > 1$, $\Omega = 1$ is the marginal case where multiple solutions exist but a small amount of *a priori* information will restore uniqueness, and for $\Omega < 1$ the problem is not unique and a multitude of solutions will exist that are consistent with the data. It is easily shown that for three-dimensional objects, $\Omega \geq 4$, and that for a three-dimensional, convex, centrosymmetric support (such as a cuboid or a three-dimensional parallelepiped), $\Omega = 4$. Therefore, the phase problem in the latter case is overdetermined by a factor four.

Note that for a real valued object, the autocorrelation is centrosymmetric and so the number of independent amplitude data is proportional to $|\mathcal{A}|/2$ and the number of object parameters is proportional to $|\mathcal{S}|$. For a complex object, the number of data is proportional to $|\mathcal{A}|$ and the number of object parameters (real and imaginary parts) is proportional to $2|\mathcal{S}|$. The constraint ratio is given by equation (1) in both cases, and there is no distinction between real and complex objects.

The constraint ratio is based on the relative number of data and parameters, so that, since the problem is nonlinear, the condition $\Omega > 1$ does not completely exclude multiple solutions in all cases. However, counter-examples exist only in contrived cases that are unlikely to arise in practice, and occur with probability zero [see Barakat & Newsam (1984) and §2 of Millane & Chen (2015) for more information].

It is useful to consider uniqueness for the phase problem for a single object in the following way. Consider, for simplicity, a real object of $M \times M \times M$ samples (or grid points), with a total of $Q = M^3$ samples. This is easily extended to other object shapes and to complex objects. Denote the sample values of the object function by $\xi_1, \xi_2, \ldots, \xi_Q$, and represent the object by the vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_Q)$ in the Q-dimensional vector space R^Q. The autocorrelation of the object is discrete with $2M \times 2M \times 2M$ samples (for large M), and has $4Q$ independent sample values (since it is centrosymmetric), and this is therefore the number of independent Fourier amplitude data. Each Fourier amplitude datum describes a constraint on $\boldsymbol{\xi}$. If we order the data, the first datum reduces $\boldsymbol{\xi}$ from belonging to R^Q to belonging to a subspace of R^Q of dimension $Q-1$. The next datum further constrains $\boldsymbol{\xi}$ to belong to a subspace of dimension $Q-2$. Continuing in this way, after Q data are included, the solution $\boldsymbol{\xi}$ is reduced to belonging to a subspace of dimension 0, *i.e.* a point set in R^Q. The structure of each subspace will be highly complex, both geometrically and topologically, but if the Q Fourier amplitude data are independent, then the dimensionality reduction will occur as described. Adding one more independent amplitude datum will select out one of the (many) points in the point set as the correct (unique) solution. Since there are $4Q$ independent data and only $Q+1$ are required, there is a data excess of $3Q-1$, and uniqueness is established. The ratio of the number of data available to that required is $4-\delta$ where δ is small and is $\mathrm{O}(Q^{-1})$. This corresponds to the constraint ratio $\Omega = 4$ for the three-dimensional cuboid case and the requirement $\Omega > 1$ for uniqueness.

3. Uniqueness for a crystal

We now extend the above ideas to the case of crystalline objects. Consider first a finite crystal (object) of $N \times N \times N$ unit cells. The volume of the object is $|\mathcal{S}_N| = N^3 V$, where $\mathcal{S}_N$ is

the support of the crystal and V is the volume of the unit cell. Since all the unit cells are the same, the number of independent object parameters is proportional to $|\mathcal{S}_N|/N^3 = V$. The normalized autocorrelation of the crystal, $A_N(\mathbf{x})$, where $\mathbf{x}$ denotes position in real space, can be written as

$$A_N(\mathbf{x}) = \frac{1}{N^3} \sum_{\mathbf{m}=-(\mathbf{N}-\mathbf{1})}^{\mathbf{N}-\mathbf{1}} (N - |m_1|)(N - |m_2|)(N - |m_3|) \times A(\mathbf{x} - \mathbf{m}\boldsymbol{\Lambda}), \quad (2)$$

where $\mathbf{m} = (m_1, m_2, m_3)$, the matrix $\boldsymbol{\Lambda} = (\mathbf{a}|\mathbf{b}|\mathbf{c})^{\mathrm{T}}$, where $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ are the unit-cell vectors, *i.e.* the rows of $\boldsymbol{\Lambda}$ are the unit-cell vectors, $\mathbf{N} = (N, N, N)$, and $A(\mathbf{x})$ is the autocorrelation of a single unit cell. This is illustrated for one dimension for $N = 3$ in Fig. 1. The volume of the support of the autocorrelation of the crystal is $|\mathcal{A}_N| = 8N^3V$. However, as a result of equation (2), not all sample values of the autocorrelation are independent. Inspection of Fig. 1, and the extension to three dimensions, shows that the whole of $A(\mathbf{x})$ can be determined from information on the boundary of $A_N(\mathbf{x})$, so that in three dimensions the volume of the autocorrelation that contains independent data is $8V$. Therefore, substituting into equation (1), the constraint ratio for the finite crystal is $\Omega_N = 8V/2V = 4$. The result is therefore the same as for a single object, as expected, and in principle the whole finite crystal could be reconstructed from a measurement of its continuous diffracted intensity. In practice, however, for all but very small crystals (small N), it would be difficult to measure the continuous diffracted amplitude between the Bragg reflections, due to its small values in these regions.

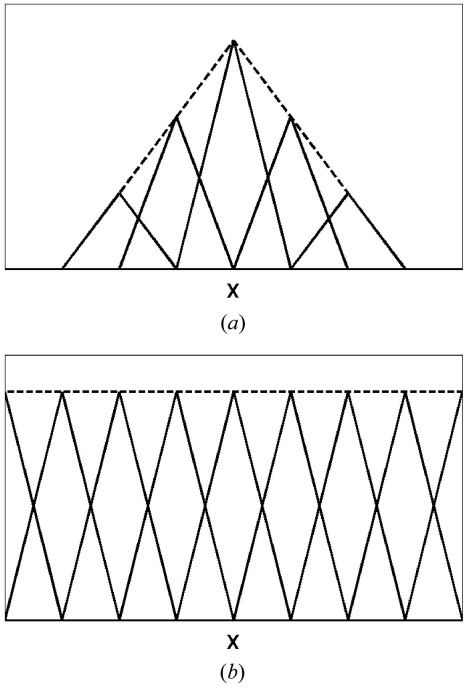


Figure 1
(*a*) The weighted autocorrelations of a single unit cell (solid lines) that make up the autocorrelation of a one-dimensional crystal with $N = 3$ unit cells (dashed line) as in equation (2). (*b*) The Patterson function (dashed line) for an infinite crystal that is made up of an infinite number of equally weighted autocorrelations of a single unit cell (solid lines).

For a realistic crystal, N is large and we have to consider the limit $N \rightarrow \infty$. The autocorrelation $A_N(\mathbf{x})$ then extends to infinity and reduces to the Patterson function $P(\mathbf{x})$, *i.e.*

$$\lim_{N\rightarrow\infty} A_N(\mathbf{x}) = \sum_{\mathbf{m}=-\infty}^{\infty} A(\mathbf{x} - \mathbf{m}\boldsymbol{\Lambda}) = P(\mathbf{x}), \quad (3)$$

which is illustrated for one dimension in Fig. 1(*b*). The boundary region of $A_N(\mathbf{x})$ is now not accessible, and all that is available is $P(\mathbf{x})$, which is periodic with a period that has volume V. Therefore, for a crystal, the number of data is proportional to V, and the constraint ratio, denoted Ω_c, is

$$\Omega_c = \frac{V}{2V} = \frac{1}{2}. \quad (4)$$

The crystallographic phase problem is therefore highly underdetermined in the absence of any additional constraints.

If additional real-space information is available, the degrees of freedom in, or the unique region of, the unit cell and the Patterson will be modified, and the constraint ratio can be written as

$$\Omega_c = \frac{|\mathcal{P}_u|}{|\mathcal{U}_u|}, \quad (5)$$

where $\mathcal{U}_u$ and $\mathcal{P}_u$ denote the unique region of the unit cell and of the Patterson, respectively. Note that the 2 in the denominator of equation (1) is now absorbed into $|\mathcal{P}_u|$ since $\mathcal{P}_u$ is always centrosymmetric. Equation (5) gives the constraint ratio for a crystal, and is a function of only the shape and symmetry of the molecule and the unit cell (since $\mathcal{U}_u$ can be calculated from this information, and $\mathcal{P}_u$ can be calculated from $\mathcal{U}_u$). The constraint ratio [equation (5)] can be used to characterize the uniqueness of the crystallographic phase problem and the effects of different kinds of real-space information.

4. Real-space constraints

Here we evaluate the constraint ratio for three kinds of real-space constraint: (i) a molecular envelope (*i.e.* the support of the molecule), (ii) crystallographic symmetry and (iii) non-crystallographic symmetry. We also consider the cases when the molecular envelope is known and when it is unknown, *a priori*.

4.1. Known molecular envelope

Consider the case where the molecule does not occupy all of the unit cell, which is essentially always the case in protein crystallography. We consider first the case where the molecular envelope is known *a priori*, and the case where it is not known is considered in the next section. The shape of the envelope can sometimes be obtained from experimental techniques such as solution scattering, electron microscopy or solvent contrast variation (Hao, 2006; Carter *et al.*, 1990; Lo *et al.*, 2009). If the shape of the molecular envelope is known, and assuming it can be positioned in the unit cell, then the

number of unknowns is proportional to its volume, *i.e.* $|\mathcal{U}_u| = pV$, where p is the fraction of the unit cell occupied by the molecule. Since a restricted molecular support (envelope) gives rise to a restricted autocorrelation support, we need to consider the possibility that the Patterson function does not occupy the whole of the unit cell, reducing the size of its unique region to less than $V/2$. Let $|\mathcal{P}_u| = qV/2$, where q denotes the proportion of the unit cell that is occupied by the Patterson, and substitution into equation (5) gives

$$\Omega_c = \frac{q}{2p}. \tag{6}$$

Since macromolecules must pack in a crystal in such a way that they make contacts with molecules in adjacent unit cells, they must occupy the unit cell in a fairly homogeneous manner. The result is that it is unlikely that the autocorrelation (of a single molecule) will not occupy all of the unit cell. It is even more unlikely that the Patterson will not occupy all of the unit cell. This is illustrated in Fig. 2. In almost all cases then, $q = 1$ and equation (6) reduces to

$$\Omega_c = \frac{1}{2p} = \frac{1}{2(1-s)}, \tag{7}$$

where s is the solvent content of the crystal. The constraint ratio then increases with increasing solvent content, as expected, and uniqueness ($\Omega_c < 1$) requires that $p < 0.5$, *i.e.* a protein content of less than 50%, or a solvent content of greater than 50%.

It is interesting to note that, since generally $q = 1$, the constraint ratio Ω_c depends, as shown by equation (7), on only the volume of the envelope, relative to that of the unit cell, and not on its shape. This is in contrast to the single object case where Ω depends on the shape of the object, rather than on its volume.

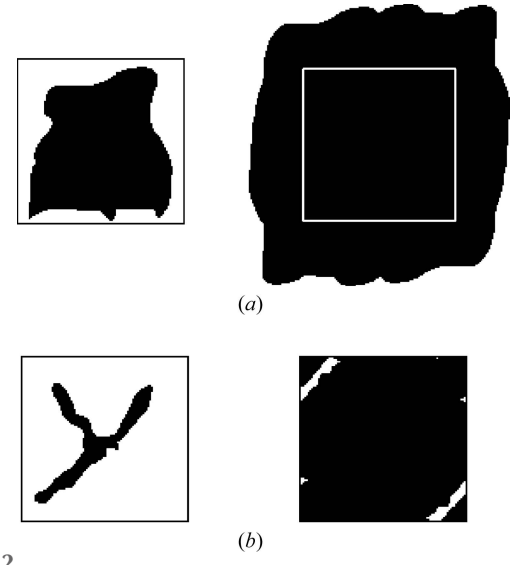


Figure 2
(*a*) A molecular envelope (left) and its corresponding autocorrelation (right) that fills the unit cell. (*b*) A molecular envelope (left) and one period of its corresponding Patterson function (right) that does not fill the unit cell. The case (*b*) requires an extremely tenuous and low-occupancy molecule, that is unlikely to occur in practice.

4.2. Unknown molecular envelope

An important caveat of the previous section is that it assumes that the molecular envelope is known. Referring to the discussion in §2, use of the number of object variables in the constraint ratio definition implicitly assumes that it is known, at least for reconstruction purposes, what those variables are. This is not the case if the envelope is unknown, since it is not known which samples (grid points) are inside the envelope. However, in many cases in protein crystallography, the envelope *shape* may be unknown, but the protein envelope (or solvent) *volume* can be estimated (Weichenberger & Rupp, 2014). Here we consider the case where the volume of the protein envelope, rather than the envelope itself, is known.

Consider a unit cell of $M \times M \times M$ samples, with a total of $Q = M^3$ samples, and known protein content p, so that the protein is known to occupy $P = pQ$ samples. The location of these P samples is unknown, however. Considering the formulation of the problem described in §2, the point is that with an unknown envelope, the molecule cannot be represented as a point $\boldsymbol{\xi}$ in R^P, because we don't know the vector-space coordinates that generate R^P. All we know is that the molecule is in the unit cell, so we need to represent it as a vector in the Q-dimensional space R^Q. The problem then is that there are only $Q/2$ amplitude data and solution to the phase problem is not unique.

However, if it is known that the object occupies only P samples, then for a particular envelope, $\boldsymbol{\xi}$ is in a P-dimensional hyperplane in R^Q (*i.e.* with the other $Q - P$ sample values fixed at zero). Furthermore, there are only a finite number of possible envelopes, *i.e.* there is a finite number of ways of selecting P samples from the Q samples. Under these conditions then, the object belongs to a P-dimensional subspace, or manifold, in R^Q, that is the union of ${}^QC_P$ P-dimensional hyperplanes (where ${}^QC_P$ denotes the number of combinations). Starting with this manifold and applying the dimensionality reduction argument in §2, the solution will again be reduced to a point set in this manifold with P data. Again, an additional datum will likely select out the correct solution from this point set. The number of independent data is $Q/2$, so uniqueness requires that $Q/2 > P = pQ$, or $p < 0.5$, *i.e.* a protein content less than 50%, or a solvent content greater than 50%. The result is therefore the same as for a known envelope, and the constraint ratio is still given by equation (7). Since the solution manifold is larger than R^P, the size of the point set may be larger than for the known envelope case, but there is still a data excess of $(\frac{1}{2} - p)Q - 1$ when $p < 1/2$. We therefore conclude that the solution to the crystallographic phase problem with only knowledge that the crystal protein content, or volume, is less than 50% of the unit cell, is also unique.

Uniqueness for the case of an unknown envelope was also investigated numerically by simulation. The idea is that since if there are multiple solutions to the problem there will be many

Table 1
Summary of simulation results.

Object size	p	Ω_c	Runs converged	Correct solutions	Average iterations for convergence
15×15	0.27	1.87	10/10	10/10	4×10^4
16×16	0.30	1.64	5/10	5/5	1×10^5
17×17	0.34	1.46	1/10	1/1	8×10^5
24×24	0.68	0.73	10/10	0/10	1×10^4

such solutions, an effective reconstruction algorithm will find one of those solutions rather easily. Iterative projection algorithms such as the difference map algorithm (Elser, 2003*a*) are effective at finding solutions to non-convex problems of this kind. By setting up such an algorithm with the appropriate constraints, the nature of the solution space can be examined by running the algorithm multiple times with different initial conditions. If multiple runs of the algorithm either converge to only the correct solution, or do not converge, then uniqueness is strongly supported. If the problem is not unique, then the algorithm will frequently converge rather quickly to an incorrect solution.

The only difficulty with this approach in the present case is that, while for the case of a known envelope the real-space constraint set (a single hyperplane) is convex, for the case of an unknown envelope the constraint set (a large number of orthogonal hyperplanes) is highly non-convex. The presence of this rather weak and highly non-convex constraint substantially increases the difficulty of the reconstruction problem, increasing the number of iterations required for convergence, potentially to an impractically large value. This necessitates simulations with small objects. On the other hand, since we are interested here in uniqueness rather than reconstruction, non-convergence is almost as informative as convergence.

We used the difference map algorithm (Elser, 2003*a*), which is an effective algorithm for phase retrieval, to study uniqueness in this way. We used a two-dimensional unit cell for convenience (the same behaviour is expected in three dimensions since in the crystallographic case, Ω_c is independent of the dimensionality). In real space, the only constraints applied are the size of the envelope (*i.e.* the number of non-zero sample values) and positivity of the electron density. In reciprocal space, the constraint is to match the structure amplitudes of the true molecule. In addition to the usual positivity and Fourier amplitude projections (Millane & Lo, 2013), the projection for the envelope size is easily shown to consist of setting the $Q - P$ smallest density values to zero and leaving the other P values unchanged, at each iteration (Elser, 2003*b*).

A 29×29 sample unit cell was used and a single square 'molecule' of various sizes was placed in the unit cell in $P1$ to vary the protein (or solvent) content, and thus vary Ω_c given by equation (7). The reconstruction algorithm was run for 10^6 iterations, starting with ten different random molecules, for each molecule size. For each run, the solution was taken as that which gives the minimum mean-square error between the resulting structure amplitudes and the data. With an unknown support in $P1$, the structure amplitudes are insensitive to the absolute position of the support, and convergence of the algorithm can be slowed by 'drifting' of the support. Therefore, the reconstruction was constrained to have its centre of mass coincident with the centre of mass of the true molecule.

The results of the simulations are summarized in Table 1. The table shows the number of runs that converged and the number of correct reconstructions for the converged runs. For the converged runs, the mean-square error in reciprocal space approached very small values. The average number of iterations required in the converged cases is also shown in the table. Convergence was obtained for $\Omega_c > 1.4$ and $\Omega_c < 0.8$ in less than 10^6 iterations. However, for $0.8 < \Omega_c < 1.4$ the algorithm did not converge within 10^6 iterations. This is due to the weak and highly non-convex real-space constraint, particularly for values of Ω_c close to unity, as mentioned above. Inspection of the table shows that in all cases for which $\Omega_c > 1$ ($p < 0.5$), the algorithm either converged to the correct solution (which therefore automatically had the correct envelope), or it did not converge. In no cases did it converge to an incorrect solution. This shows strong support for uniqueness in the case $\Omega_c > 1$. For $\Omega_c = 0.73$, multiple incorrect solutions were easily found by the algorithm. This indicates that, indeed, non-unique solutions are likely to be found if they exist.

It is interesting to consider the number of hyperplanes in the solution set. For each envelope shape there are Q possible positions of the envelope (including those that wrap around the unit-cell edges), and these should be treated as redundant since they all give the same Fourier amplitude. Therefore, the number of envelope-position-independent hyperplanes, denoted $\mathcal{N}_h(p, Q)$, is

$$\mathcal{N}_h(p, Q) = Q^{-1}\, {}^{Q}C_{pQ}. \tag{8}$$

For fixed Q, this quantity is a maximum at $p = 0.5$. For typical protein crystal solvent contents between 70 and 30% (*i.e.* $0.3 < p < 0.7$), which represents 95% of the entries in the PDB, the dependence of $\mathcal{N}_h(p, Q)$ on p is weak, and $\mathcal{N}_h(p, Q)$ is given approximately by (see Appendix *A*)

$$\mathcal{N}_h(p, Q) \simeq Q^{-3/2}\, 2^Q, \qquad p \simeq 0.5. \tag{9}$$

The number of hyperplanes is therefore indeed large, but the simulations show that the structure amplitude data are able to select out the correct hyperplane corresponding to the solution, in spite of this large number. For example, although the cases described above are for small objects, the number of variables is about 10^3, and the number of hyperplanes is about 10^{250}. This number of hyperplanes emphasizes the extreme non-convexity of the real-space constraint set.

The algorithm described above is useful for investigating uniqueness, but it is not a practical approach in protein crystallography where the number of sample values is much larger and many more iterations would be required. However, in practice, more is known about protein envelopes. In particular, protein envelopes are generally quite compact. This property substantially reduces the number of possible envelopes and the number of hyperplanes, significantly easing the recon-

struction problem. Supplementing the reconstruction algorithm with additional compactness constraints through the use of, for example, smoothing and shrinking of the support (Wang, 1985; Marchesini *et al.*, 2003) or other schemes (Lo *et al.*, 2009), should allow *ab initio* phasing without initial envelope information for practical problems. Indeed, the recent results of He & Su (2015) support this conclusion.

In summary then, even in cases where the molecular envelope is not known *a priori*, the macromolecular crystallographic phase problem has a unique solution if the protein content of the crystal is less than 50%.

4.3. Crystallographic symmetry

Consider now the effect of crystallographic (space-group) symmetry on the constraint ratio. For non-centric crystallographic symmetry of order R, the Patterson function has symmetry of order $2R$ (as illustrated in Fig. 3*a*). We then have that $|\mathcal{U}_u| = V/R$ and $|\mathcal{P}_u| = V/2R$, and substitution into equation (5) gives

$$\Omega_c = 1/2, \tag{10}$$

i.e. the same as for the case without symmetry. For centric crystallographic symmetry of order R, the Patterson has symmetry of order R (as illustrated in Fig. 3*b*). In this case, $|\mathcal{P}_u| = V/R$, and substitution into equation (5) gives

$$\Omega_c = 1. \tag{11}$$

This is then the marginal case that corresponds to a countable number of phase solutions (*i.e.* two choices for each reflection) and only a small amount of additional *a priori* information is required to render the solution unique. These results are consistent with the well known fact that reduction in the number of parameters due to the crystallographic symmetry is exactly matched by the same number of relationships between the structure amplitudes, and the overall data/parameter ratio remains unchanged. Crystallographic symmetry does not therefore constrain the phase problem, except in the centric case which does not occur with biomolecules.

4.4. Non-crystallographic symmetry

Consider now NCS of order R. NCS does not lead to increased symmetry in the Patterson function (see the illustration in Fig. 3*c*), so that $|\mathcal{U}_u| = V/R$ and $|\mathcal{P}_u| = V/2$, and substitution into equation (5) gives

$$\Omega_c = R/2. \tag{12}$$

The redundancy of the phase problem is therefore improved by a factor R, and a unique solution is expected in principle if $R > 2$. Therefore, as a result of equation (12), NCS is a significant factor for *ab initio* phasing. This result coincides with early considerations of the effect of NCS on constraining the phases (Crowther, 1969; Bricogne, 1974), and is related to the fact that NCS, unlike crystallographic symmetry, does not lead to relationships between the structure-factor amplitudes, and so the number of independent data is not reduced. An alternative interpretation is that R-fold NCS leads to a denser sampling, by a factor R, relative to the Bragg sampling, of the continuous Fourier amplitude of the contents of the unit cell, increasing the number of data by a factor R (Millane, 1990, 1993).

As with the case of a known molecular envelope, the above analysis assumes that the NCS operators are known (so that the number of electron-density parameters can be reduced by a factor R). This problem is not so difficult, however, as the order of the NCS can be determined from a self-rotation function (Tong & Rossmann, 1997), although positioning of the NCS origin in the unit cell can present difficulties.

NCS is always accompanied by a restricted molecular envelope, and combining the above results gives

$$\Omega_c = \frac{R}{2p} \tag{13}$$

in the presence of both constraints. Therefore, with both constraints, solution to the phase problem is expected to be considerably eased. For example, with twofold NCS and 50% solvent content, or with threefold NCS and 25% solvent content, $\Omega_c = 2$ and the problem is expected to be well determined in practice.

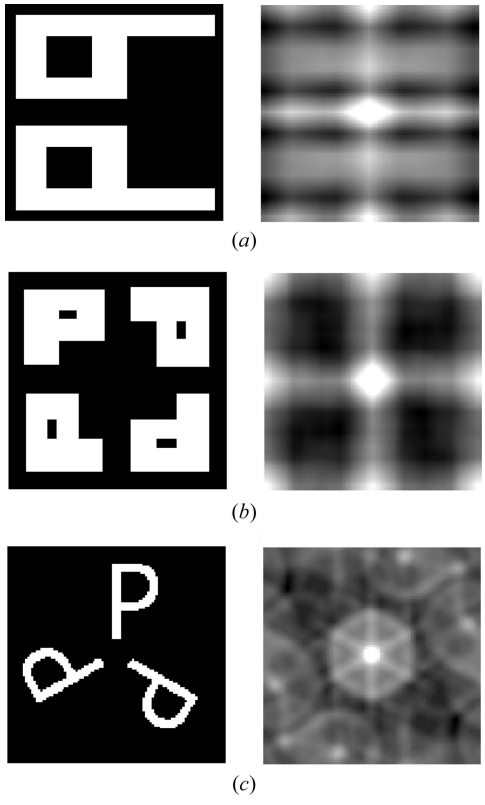


Figure 3
Examples of two-dimensional unit cells (left) and one period of their corresponding Patterson functions (right) for (*a*) non-centric *pm* crystallographic symmetry, (*b*) centric *p*4 crystallographic symmetry and (*c*) non-crystallographic threefold symmetry in plane group *P*1, as described in the text. The corresponding Patterson symmetries are (*a*) *p*2*mm*, (*b*) *p*4 and (*c*) *p*2.

5. Summary

The constraining power of real-space information in protein crystallography is conveniently characterized by a constraint ratio that can be calculated using equation (5). The constraint ratio is useful in that it gives guidance on the likely success of *ab initio* phasing. For example, recent results indicate that, as a result of errors and missing data, a value of Ω greater than about 1.5 might be needed for *ab initio* phasing in practice (Liu *et al.*, 2012; Millane & Lo, 2013). Equation (5) allows the constraint value to be calculated for specific kinds of real-space information in order to make this assessment.

For the case of protein content and NCS, the constraint ratio is given by equation (13). Evaluation of this equation suggests that, with the use of suitable reconstruction algorithms, *ab initio* phasing should be feasible with quite modest values of these parameters. Recent results using iterative projection algorithms indicate that this is the case (Liu *et al.*, 2012; He & Su, 2015; Lo *et al.*, 2015). NCS is a particularly powerful constraint if incorporated into iterative projection algorithms (Millane & Lo, 2013; Lo *et al.*, 2015).

Although an estimate of the molecular envelope is desirable if available, uniqueness does not depend on *a priori* knowledge of the envelope, and envelope volume and compactness are a powerful constraint. The recent results of He & Su (2015) support this conclusion. Overall, these results indicate that more comprehensive tests of the application of iterative projection algorithms to phase retrieval in protein crystallography are warranted.

APPENDIX *A*

The number of envelope-position-independent hyperplanes $\mathcal{N}_h(p, Q)$ in R^Q is given by equation (8). Since Q is large, applying Stirling's approximation to ${}^{Q}C_{pQ}$ gives

$$\mathcal{N}_h(p, Q) \simeq \left(\frac{2}{\pi}\right)^{1/2} \frac{1}{2[p(1-p)]^{1/2}} Q^{-3/2}\left[p^{-p}(1-p)^{p-1}\right]^{Q}. \tag{14}$$

For fixed Q, $\mathcal{N}_h(p, Q)$ is symmetric about $p = 0.5$, where it is a maximum. At $p = 0.5$, equation (14) reduces to

$$\mathcal{N}_h(0.5, Q) \simeq \left(\frac{2}{\pi}\right)^{1/2} Q^{-3/2} 2^{Q}. \tag{15}$$

For p close to 0.5, the dependence on p is fairly weak. For example, for $p = 0.3$, substitution into equation (14) gives

$$\mathcal{N}_h(0.3, Q) \simeq \left(\frac{2}{\pi}\right)^{1/2} (1.09) Q^{-3/2} (1.84)^{Q}. \tag{16}$$

Therefore, noting that $(2/\pi)^{1/2} \simeq 0.8$, for $0.3 < p < 0.7$, $\mathcal{N}_h(p, Q)$ can be suitably approximated by

$$\mathcal{N}_h(p, Q) \simeq Q^{-3/2} 2^{Q}, \qquad 0.3 < p < 0.7. \tag{17}$$

Acknowledgements

This work was supported by a James Cook Research Fellowship and a Marsden grant to RPM, and a University of Canterbury College of Engineering Doctoral Scholarship to RDA.

References

Barakat, R. & Newsam, G. (1984). *J. Math. Phys.* **25**, 3190–3193.
Bates, R. H. T. (1984). *Comput. Vision Graph. Image Process.* **25**, 205–217.
Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.
Bruck, Y. M. & Sodin, L. G. (1979). *Opt. Commun.* **30**, 304–308.
Carter, C. W., Crumley, K. V., Coleman, D. E., Hage, F. & Bricogne, G. (1990). *Acta Cryst.* A**46**, 57–68.
Crowther, R. A. (1969). *Acta Cryst.* B**25**, 2571–2580.
Drenth, S. (1994). *Principles of Protein X-ray Crystallography*. New York: Springer-Verlag.
Elser, V. (2003*a*). *J. Opt. Soc. Am. A*, **20**, 40–55.
Elser, V. (2003*b*). *Acta Cryst.* A**59**, 201–209.
Elser, V. & Millane, R. P. (2008). *Acta Cryst.* A**64**, 273–279.
Fienup, J. R. (1978). *Opt. Lett.* **3**, 27–29.
Hao, Q. (2006). *Acta Cryst.* D**62**, 909–914.
He, H. & Su, W.-P. (2015). *Acta Cryst.* A**71**, 92–98.
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. L., Kingston, R. L. & Millane, R. P. (2015). *Acta Cryst.* A**71**, 451–459.
Marchesini, S. (2007). *Rev. Sci. Instrum.* **78**, 011301.
Marchesini, S., He, H., Chapman, H. N., Hau-Riege, S. P., Noy, A., Howells, M. R., Weierstall, U. & Spence, J. C. H. (2003). *Phys. Rev. B*, **68**, 140101.
Miao, J., Sayre, D. & Chapman, H. N. (1998). *J. Opt. Soc. Am. A*, **15**, 1662–1669.
Millán, C., Sammito, M. & Usón, I. (2015). *IUCrJ*, **2**, 95–105.
Millane, R. P. (1990). *J. Opt. Soc. Am. A*, **7**, 394–411.
Millane, R. P. (1993). *J. Opt. Soc. Am. A*, **10**, 1037–1045.
Millane, R. P. (1996). *J. Opt. Soc. Am. A*, **13**, 725–734.
Millane, R. P. & Chen, J. P. J. (2015). *J. Opt. Soc. Am. A*, **32**, 1317–1329.
Millane, R. P. & Lo, V. L. (2013). *Acta Cryst.* A**69**, 517–527.
Sheldrick, G. M., Gilmore, C. J., Hauptman, H. A., Weeks, C. M., Miller, R. & Uson, I. (2011). *International Tables for Crystallography*, Vol. F., edited by E. Arnold, D. M. Himmel & M. G. Rossmann, pp. 413–429. Chichester: Wiley.
Tong, L. & Rossmann, M. G. (1997). *Methods Enzymol.* **276**, 594–611.
Wang, B. C. (1985). *Methods Enzymol.* **115B**, 90–112.
Weichenberger, C. X. & Rupp, B. (2014). *Acta Cryst.* D**70**, 1579–1588.