Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 15 September 2014
Accepted 31 October 2014

Direct phasing of protein crystals with high solvent content

Hongxing He and Wu-Pei Su*

Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, Texas 77204, USA. Correspondence e-mail: wpsu@uh.edu

An iterative transform method is proposed for solving the phase problem in protein crystallography. In each iteration, a weighted average electron-density map is constructed to define an estimated protein mask. Solvent flattening is then imposed through the hybrid input–output algorithm [Fienup (1982). *Appl. Opt.* **21**, 2758–2769]. Starting from random initial phases, after thousands of iterations the mask evolves into the correct shape and the phases converge to the correct values with an average error of 30–40° for high-resolution data for several protein crystals with high solvent content. With the use of non-crystallographic symmetry, the method could potentially be extended to phase protein crystals with less than 50% solvent fraction. The new phasing algorithm can supplement and enhance the traditional refinement tools.

1. Introduction

Finding the phases of diffracted X-rays is an important step in protein structure determination. Although the use of selenomethionine and multiple anomalous dispersion (MAD) has rendered the procedure almost routine, the time and resources involved can still be substantial for many large proteins, not to mention the difficulty of expressing some selenomethionine-substituted proteins in eukaryotic hosts (Strop *et al.*, 2007). Alternate techniques that reduce the experimental and investigator demands are therefore still of considerable importance.

Recently, an iterative transform algorithm has been proposed by Liu *et al.* (2012) to retrieve the phases. An envelope of the region occupied by the protein inside the unit cell is assumed. In each iteration, Fourier refinement (replacing calculated Fourier amplitudes by observed ones) is combined with a density modification in real space, which is essentially a gradual solvent flattening through the hybrid input–output (HIO) algorithm (Fienup, 1982). For a peculiar choice of the protein boundary, Liu *et al.* were able to recover several high-resolution structures with high solvent content. As such, that work constitutes important progress in solving the phase problem. It is obviously desirable to eliminate the requirement of a prior knowledge of the protein region. As will be demonstrated below, it is possible to do so and therefore possible to directly phase protein crystals with high solvent content.

We have followed basically Liu *et al.*'s algorithm except that we allow the protein boundary to evolve with iteration. In each iteration, a weighted average density map is constructed to define the protein region. Thus the protein boundary is not assumed beforehand, rather it is dynamic and becomes accurate only at the end of successful calculations. Therefore, our procedure is *ab initio* phasing. A very similar idea has been pursued by Millane & Stroud (1997) and by van der Plas & Millane (2000) in their reconstruction of icosahedra virus images from Fourier intensities. Also the related idea of a dynamic support has been studied in the field of coherent diffraction imaging (Marchesini *et al.*, 2003; Dronyak *et al.*, 2009).

Although our primary interest is in *ab initio* phasing, our method can also be used for phase extension. Prior knowledge of low-resolution phases (say 10 Å) leads to fast convergence of high-resolution structures.

This paper is organized as follows. In §2 our methodology is described. In §3 results of trial calculations are reported. The possibility of direct phasing for low-solvent-content crystals with non-crystallographic symmetry (NCS) is examined in §4. The paper ends with a discussion in §5 and a conclusion.

2. Methodology

Our iterative transform algorithm is represented by the flowchart in Fig. 1. Each iteration begins with a real-space density g^m on the upper-left corner from the previous round of calculation. A fast Fourier transform of the density yields phases which are combined with the observed Fourier magnitudes to produce (*via* a fast inverse Fourier transform) the new electron density ρ^{m+1}. On each grid point in the unit cell, a weighted average density can be derived from the electron density ρ_j as follows:

$$W_i = \sum_{j \text{ for } d_{ij}<r} \frac{r - d_{ij}}{r} \rho_j, \tag{1}$$

where the subscript i or j represents a grid point in the unit cell. d_{ij} is the distance between two grid points. r is a radius

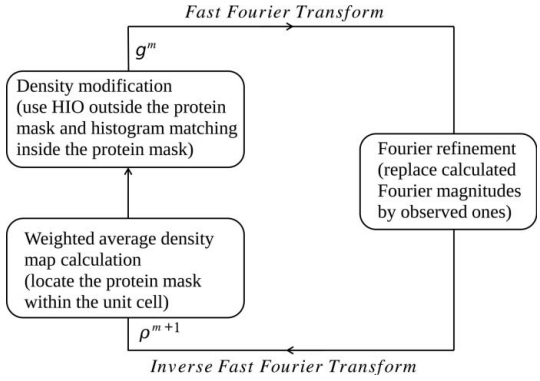

Figure 1
A flowchart of the iterative transform algorithm described in §2.

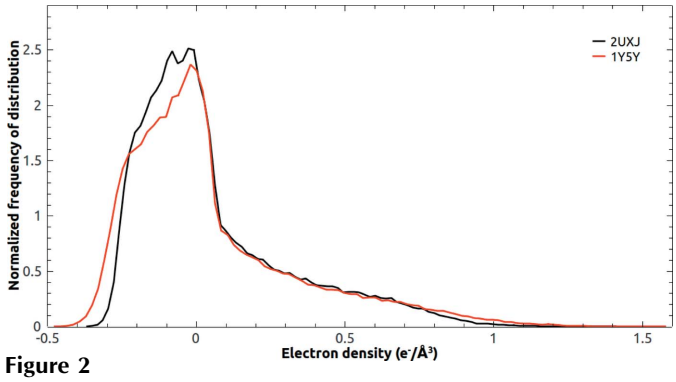

Figure 2
Density histogram inside a loose protein mask for 2uxj (black) and 1y5y (red) at 2.25 Å resolution.

which can be used to control the convergence of the solvent region. Initially, r can be as big as half of the unit-cell dimension. During the iterations, r is gradually reduced to an appropriate value.

We can also use a Gaussian function to calculate the weighted average map:

$$W_i = \sum_j \exp[-d_{ij}^2/(2\sigma^2)]\rho_j, \qquad (2)$$

where σ again can be used to control the convergence of the solvent region.

The weighted average defined in equations (1) or (2) can be easily evaluated using the convolution theorem. For a given solvent content, a proper cutoff value $W_{\rm cutoff}$ can be found by adjusting it such that the calculated solvent content agrees with the expected value. A grid point is regarded to be inside the mask if and only if the average density there is above $W_{\rm cutoff}$.

Outside the protein mask, the density is gradually driven to zero through the hybrid input–output algorithm:

$$g_i^{m+1} = \begin{cases} \rho_i^{m+1} & \text{if } W_i \geq W_{\rm cutoff} \\ g_i^m - \varepsilon\rho_i^{m+1} & \text{otherwise,} \end{cases} \qquad (3)$$

where the superscript labels the iteration number and the subscript represents a grid point in the unit cell. ε is a feedback parameter which can be used to optimize the convergence of the algorithm. ρ_i and g_i correspond to electron density before and after the HIO density modification and histogram matching. The next round of iteration begins with the new density g_i^{m+1}.

During the iterations, the estimated protein mask is dynamically updated and gradually converges to the correct shape. Histogram matching is used to modify the electron density inside the mask so that the correct electron density of the molecule emerges progressively.

3. Trial calculations

Following Liu *et al.*, we will focus on trial calculations carried out for a photosynthetic reaction center structure (PDB code 2uxj, Koepke *et al.*, 2007). The space group is $P4_32_12$. The cell dimensions are a = 139.376, b = 139.376 and c = 235.041 Å. There are 7707 non-hydrogen protein atoms in the asymmetric unit and the solvent content is 76.56%. The crystal diffracts to 2.25 Å, with lowest resolution at 27.12 Å. The completeness of the data set is 94%, with a total R value of 0.195. It is a good data set, but just like any typical set, there are reflections missing including the 92 reflections below 27.12 Å. The magnitudes of missing reflections below 2.25 Å including F_{000} are automatically reconstructed during the iterations. In other words, the intensities of missing reflections are calculated from the estimated density function in each iteration. The unit cell is discretized for fast Fourier transform, the distance between two nearby grid points is 1 Å. The density function is defined only on the grid points.

In principle, we can choose a tight mask for the protein region (*i.e.* 23% of the unit-cell volume) during each iteration step. But we have found that a somewhat different choice seems to be more effective. We have chosen a loose mask which includes 31% of the unit-cell volume, *i.e.* the volume of the protein plus 8% solvent. Correspondingly, the solvent region computed from the average density map occupies 69% of the unit-cell volume. This choice is motivated by the thinking that during the iterations, the computed boundary might not match the surface of the protein tightly. With that choice, the density histogram inside the protein mask is shown as the black curve in Fig. 2. The corresponding density histogram of a somewhat smaller formaldehyde-activating enzyme (Fae) structure [PDB code 1y5y (Acharya *et al.*, 2005), in the same space group, with a smaller solvent fraction 68%] is shown as the red curve in the same figure. There is a

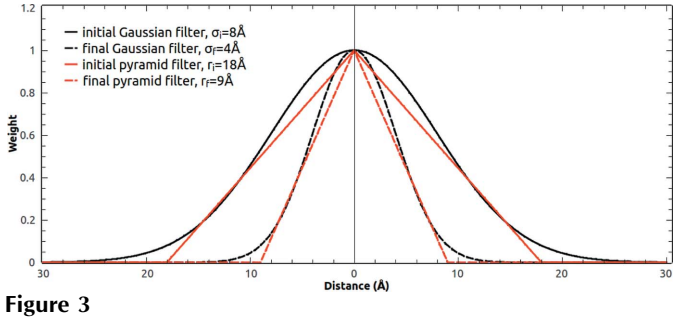

Figure 3
Weighting functions [equations (1) and (2)] used for calculating the average density at the beginning and end of an iterative run.

substantial difference between the two histograms. It turns out that they lead to the same result. As a note, the density histograms are calculated in the standard way. In the calculation of the reference density histogram, a choice of F_{000} is made so that the average density inside the protein mask is about 0.05 e Å^{-3}. This choice was found to work empirically. It partially reflects the solvent contained within the protein mask.

To monitor the evolution of the iteration, we compute the mean error in phase angle ($\Delta\varphi$) defined as follows:

$$\Delta\varphi = \frac{\sum_{h,k,l} \arccos\left\{\cos\left[\varphi^{\rm true}(h,k,l) - \varphi^{\rm cal}(h,k,l)\right]\right\}}{\sum_{h,k,l} 1}. \quad (4)$$

After some trial and error, a good choice of the width (σ) of the Gaussian weighting function in equation (2) is found to decrease linearly from 8 to 4 Å in 10 000 iterations. Alternatively, a pyramidal weighting function in equation (1) can be used for which the radius r decreases from 18 to 9 Å in the same number of iterations, as shown in Fig. 3. The HIO feedback parameter in equation (3) is taken to be $\varepsilon = 0.9$.

With the above choice of parameters, a batch of 20 independent calculations (with different random starting phases) are carried out. The phase error of the eight successful runs and one unsuccessful run is depicted in Fig. 4(*a*). In all the successful runs, the error drops suddenly from 90° to about 50°. After 8000 iterations, the HIO scheme is gradually turned off and a complete solvent flattening is imposed after 9500 iterations. That leads to a further drop of the phase error to about 32°, whereas for the unsuccessful run, the phase error remains at 90° (random phases).

It is also very instructive to examine the evolution of the R value, which is calculated from the density function g_i^{m+1} defined earlier in connection with the flowchart. As is well known, $R_{\rm free}$ is the R value that correlates well with the phase error (Brünger, 1992). For the $R_{\rm free}$ calculation, 5% of the diffraction intensity data is set aside from the working set. For the same runs depicted in Fig. 4(*a*), their $R_{\rm free}$ values are shown in Fig. 4(*b*). Clearly, the $R_{\rm free}$ value tracks the phase error in the sudden drop. After solvent flattening, the successful runs all end up with a unique $R_{\rm free}$ value of about 0.23. The $R_{\rm work}$ value is actually quite informative too, as shown in Fig. 4(*c*). Since the sudden drop of the R value is a good indicator of a corresponding improvement in phase error, it is very useful for a new structure determination.

A phase error of 32° means the final density map is very accurate. Examples of 2.25 Å maps are shown in Fig. 5. Some of the water molecules are visible in the map. It is also of interest to have a look at the protein mask near the end of a successful run (Figs. 6 and 7). There are still small parts of the protein sticking out of the mask despite the accuracy of the calculated phases.

As an example of our ability to invert lower-resolution data, we cutoff the diffraction data of 2uxj at 3.5 Å resolution and used them in a calculation. A loose protein mask which occupies 31% of the unit cell was updated in iterations. The reference density histogram was computed from 1y5y at 3.5 Å resolution. The evolution of the phase error and $R_{\rm free}$ is shown in Fig. 8. Among a batch of 20 runs, five are successful with a mean phase error of 34°. Typical electron-density maps are shown in Fig. 9.

Aside from 2uxj, we have examined several other structures with good diffraction data and high solvent content, including 1y5y and a human thyroid hormone receptor (3ilz, Bleicher *et*

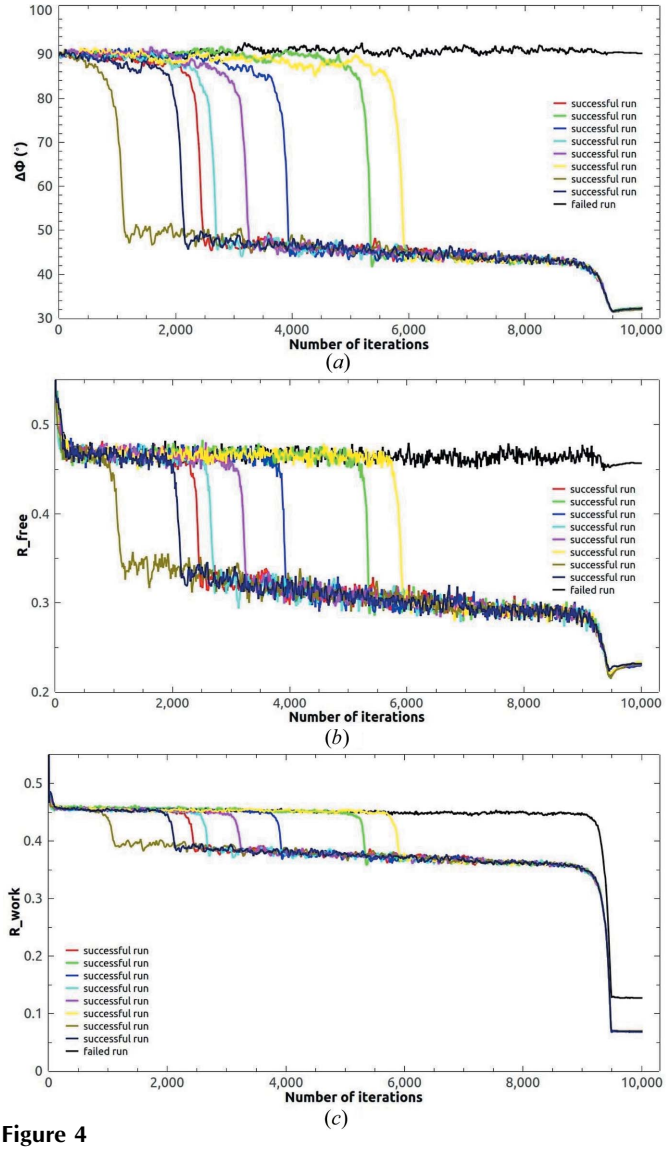

Figure 4
Evolution of the phase error and R values for eight successful and one unsuccessful runs of 2uxj described in §3.

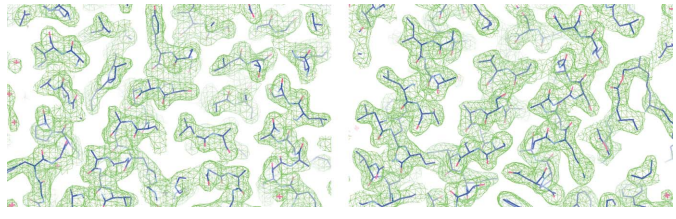

Figure 5
Two calculated 2.25 Å electron-density maps of 2uxj from successful runs in Fig. 4.

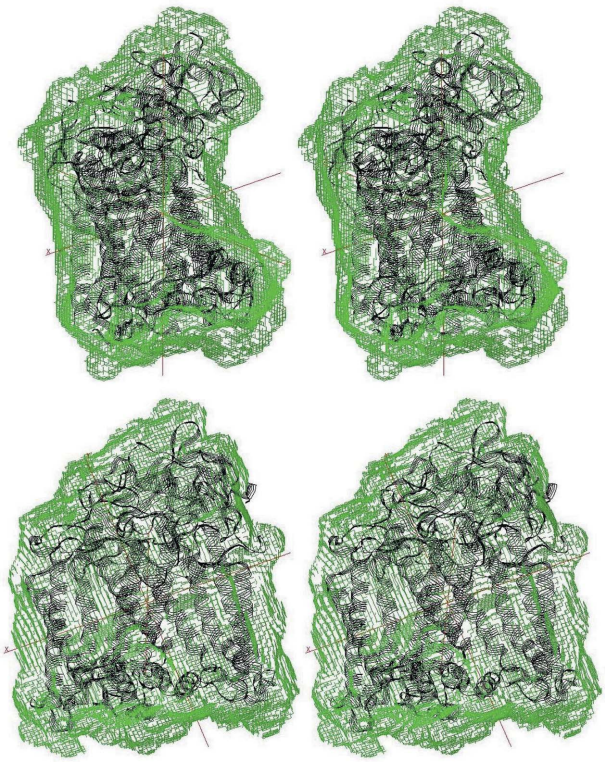

Figure 6
Stereograms of calculated protein mask compared with an atomic model of 2uxj in strands.

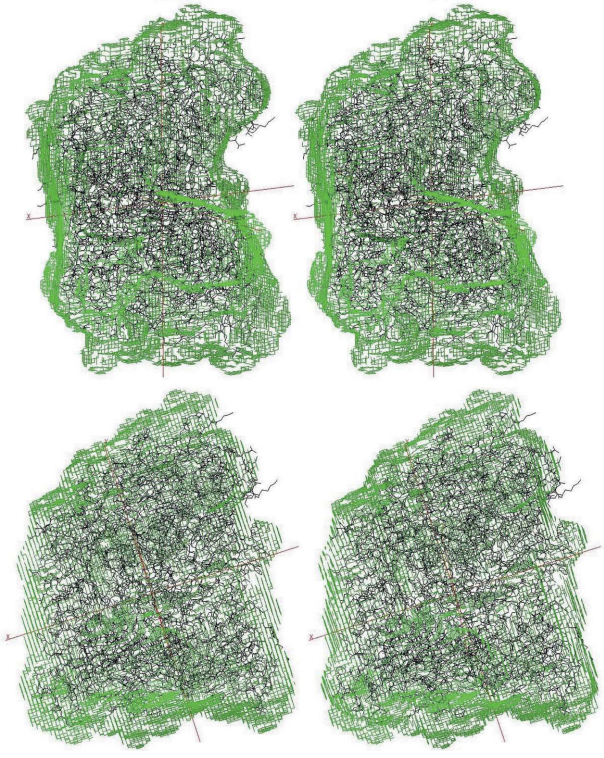

Figure 7
Stereograms of calculated protein mask compared with an atomic model of 2uxj in wireframe.

al., 2008). The last one is in the $P2_12_12_1$ space group. The calculated results are similar to those of 2uxj. In particular, the average phase errors are around 30–40°. The Fae (1y5y) has a fivefold non-crystallographic symmetry, which was not used in the calculations. Since we have used the density histogram of 1y5y to retrieve the phases of 2uxj, one might expect that the histogram of 2uxj can be used to retrieve the phases of 1y5y. It turns out to be untrue for 2 Å data. Instead, the histogram of 1ejb (Meining *et al.*, 2000) works. The evolution of phase error is displayed in Fig. 10. More than 10 000 iterations are needed for some runs to converge. The final average phase error is about 38° for 2 Å data. Typical electron-density maps are shown in Fig. 11 and the final protein mask is depicted in Fig. 12. It should be noted that a phase error of 30° can be achieved with the histogram of 1y5y itself. Therefore there are probably other structures whose histograms match 1y5y better than 1ejb. Another note is that for 1y5y, a tight protein mask actually works better than a loose mask in terms of success rate. Finally, at 2.25 Å the histogram of 2uxj can indeed be used to retrieve the phases of 1y5y.

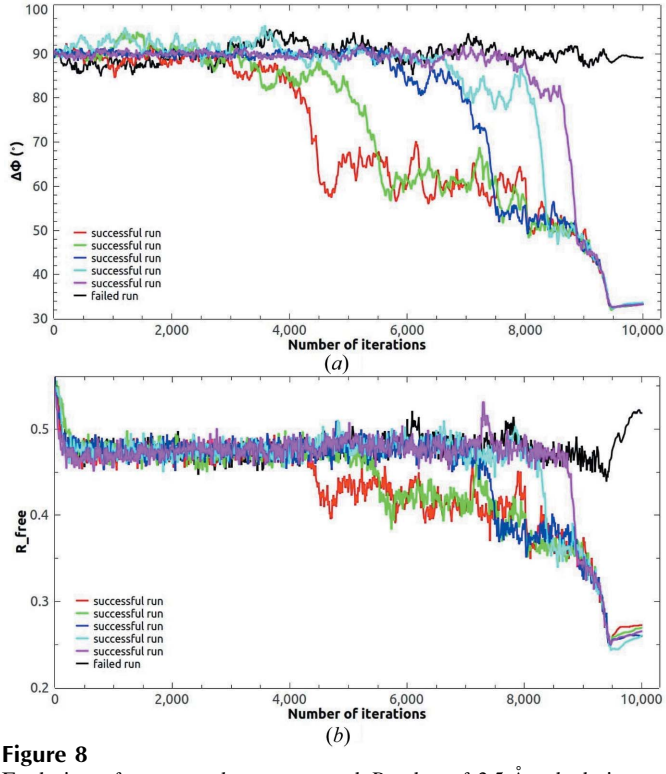

Figure 8
Evolution of average phase error and R value of 3.5 Å calculations of 2uxj.

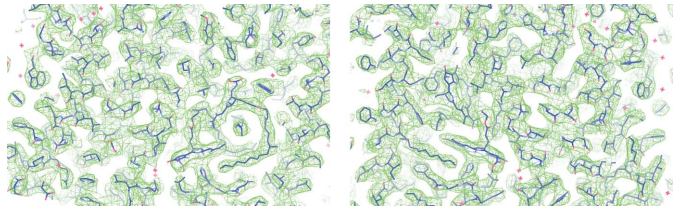

Figure 9
Typical calculated electron-density maps of 2uxj at 3.5 Å resolution.

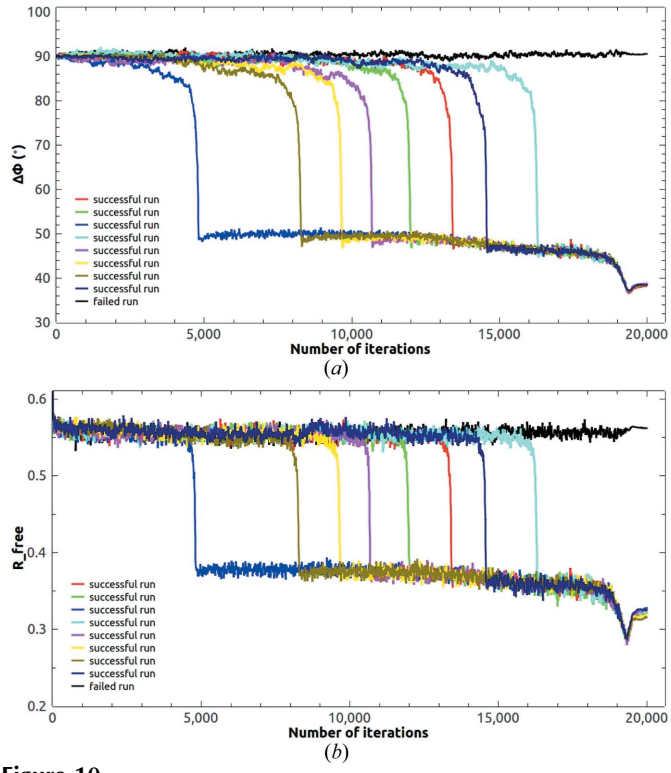

Figure 10
Evolution of phase error and R value of 2 Å runs of 1y5y described in §3.

4. Non-crystallographic symmetry

An important criterion for iterative algorithms such as HIO to work is the requirement that the number of independently measured data points exceeds the number of unknown variables, as first pointed out by Miao *et al.* (1998). Thus it is not surprising that with the use of non-crystallographic symmetry (NCS), our method may be extended to phase protein crystals with less than 50% solvent fraction as the NCS reduces that number of unknown variables (the electron density within the protein mask). Liu (2012) has illustrated that using an artificial structure possessing NCS, assuming that the envelope is available. Millane & Lo (2013) have emphasized the same point. To further demonstrate that possibility, we have studied the structure of a carbamoyltransferase (PDB code 4nf2, Center for Structural Genomics of Infectious Diseases, unpublished work). The space group is $P2_12_12_1$. The cell dimensions are $a = 85.89$, $b = 99.89$ and $c = 118.99$ Å. The NCS axis is threefold. The solvent fraction is 44.79%. The resolution range of the diffraction data extends from 29.23 to 1.74 Å, with an R value of 0.147.

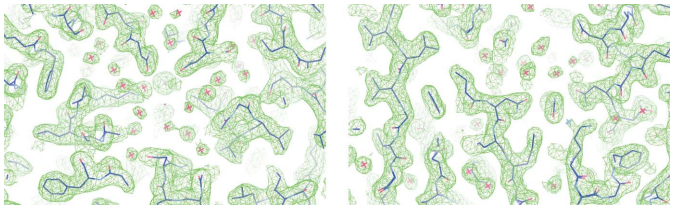

Figure 11
Typical calculated electron-density maps of 1y5y at 2 Å resolution from successful runs in Fig. 10.

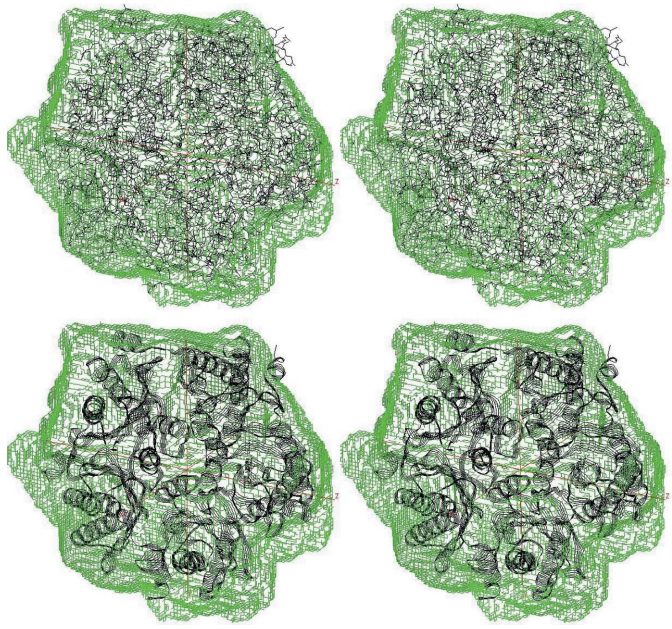

Figure 12
Stereograms of calculated protein mask for 1y5y from successful runs in Fig. 10.

As a first step in showing the possibility of an iterative phasing scheme, we assume a given low-resolution envelope (calculated from phases at 30 Å resolution) and the orientation and position of the threefold axis. Synthetic diffraction data are used instead of real data, but bulk solvent correction is taken into account. With those we have carried out HIO iterations with the 1.74 Å data. A reference density histogram was computed from 4nf2 itself at 1.74 Å resolution. Starting essentially from random phases, the evolutions of the phase error at three resolution levels are shown in Fig. 13. The protein mask is kept fixed throughout the iterations, and the threefold symmetry inside the protein envelope is enforced by conventional NCS averaging in updating the density function. It is clear from Fig. 13 that correct phases can be retrieved after many iterations despite the low solvent content.

While the orientation of the NCS axis can in general be calculated from the self-rotation Patterson correlation function, the position of the axis and the protein mask to be used for threefold rotations are unknown before the crystal is phased. It is not clear how to calculate those entities iteratively, unlike the solvent boundary. Before this problem is solved, it is useful to consider the possibility of obtaining the information about the NCS axis and protein envelope from another independent approach – a real-space approach to the phase problem that we have been exploring.

The real-space approach is a scheme of generating approximate density maps directly from the diffraction intensity data. The unit cell is discretized into grid points. On each grid point, one can assign one or zero corresponding to a point scatterer or none. For any given configuration of the

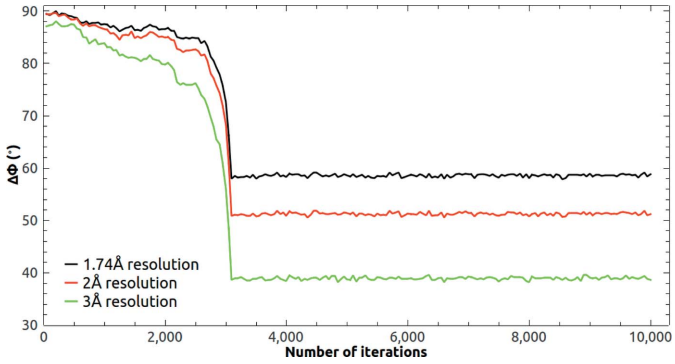

Figure 13
Evolution of phase error of several resolution ranges of 4nf2 with a fixed protein mask described in §4.

lattice (the pattern of zero's and one's) and a selected diffraction data set, the R value can be calculated. By flipping the zero and one into each other on each grid point, one can optimize the R value of the configuration. In essence, this is a construction of optimal binary maps directly from the diffraction intensities. More details can be found in Su (2008), where examples of protein envelopes constructed from real diffraction data can be found.

For 4nf2, we have constructed density maps directly from the 6–8 Å experimental diffraction data. Those maps are compared with the atomic model to show that they can provide the information we are looking for. In Fig. 14, a binary map calculated with 8 Å data (indicated by the green dots) is superimposed on the atomic model in wireframe. It is a slice of the unit cell perpendicular to the NCS axis. From the stereograms, three copies of the proteins can be seen contained approximately within an inverted triangular boundary. The NCS axis runs through the center of the inverted triangle. A channel containing this axis is actually visible in a composite density map from two (the green and red dots) independent 6 Å calculations, as shown in Fig. 15. There are $36 \times 36 \times 44$ grid points in the unit cell in both 6 Å and 8 Å calculations.

5. Discussion

For protein crystals with 50% or higher solvent content, Liu *et al.* have provided strong evidence that phasing through iterative transform is possible provided a protein envelope is available. What we have shown through the trial calculations of 2uxj and other structures is that the assumption of an envelope is not necessary, and therefore *ab initio* phasing is possible. While further trial calculations involving more types of structures remain to be carried out, the generality of our methodology strongly supports the claim that direct phasing is possible for many high-solvent-content protein crystals.

Many protein crystals with lower solvent content possess NCS. We have used 4nf2 as an example to argue that for those crystals a real-space approach may yield enough information about the NCS axis and the protein mask so that an iterative scheme can be employed to find accurate phases.

Although our primary concern in this paper is direct phasing, the algorithm can be used to supplement and enhance

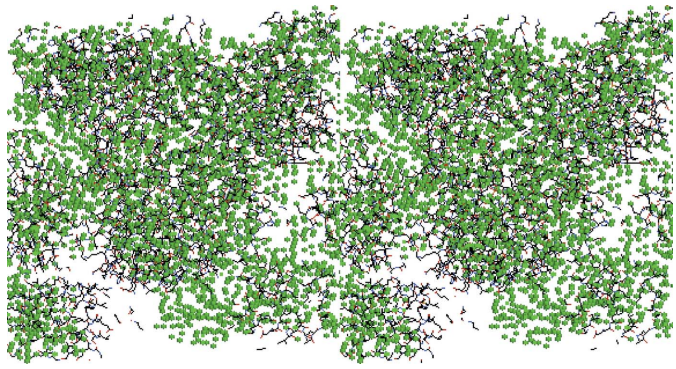

Figure 14
Binary density (stereo) map calculated from 8 Å data of 4nf2 compared with the atomic model in wireframe.

many existing refinement tools. Partial knowledge of some of the phases, for example, can easily be incorporated in the iteration. Prior real-space information such as solvent region (Zou & Phillips, 1998; Lo *et al.*, 2009) or protein fragments can also be employed in the density modification. The large number of iterations helps to eliminate the bias of initial phases.

Although our *ab initio* iterative phasing algorithm resembles the conventional solvent flattening (Wang, 1985) refinement, there are important differences. It is instructive to make a comparison. First, the density modification in the solvent region *via* the HIO scheme is much more powerful than simply setting the solvent density to a constant. It is well known, for example, that HIO can overcome the stagnation problem and therefore it makes the convergence toward the correct solution possible, whatever the initial starting phases are. Secondly, the number of iterations matters. We have seen in the trial calculations that tens of thousands of iterations are needed in general to retrieve the phases correctly. Thirdly, the missing low-resolution central reflections (due to the beam stop) are not included in the conventional refinement, but they

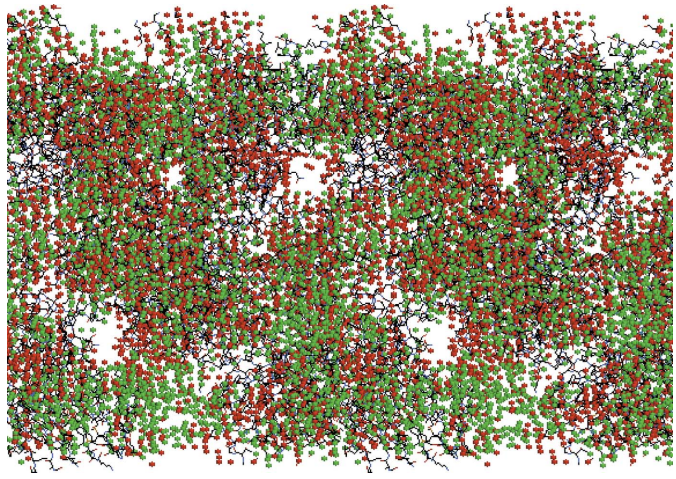

Figure 15
Binary density (stereo) maps calculated from 6 Å data (red and green dots) of 4nf2 exhibiting a channel along the NCS axis.

are dominant terms in the Fourier expansion of the density function and they can greatly affect the construction of the shape (Miao & Sayre, 2000; Miao *et al.*, 2000) of the protein. In our phasing scheme, they are reconstructed (Chen *et al.*, 2007; Liu *et al.*, 2012) from other existing reflections; thus a very complete Fourier expansion of the density is achieved, and therefore very accurate phases can be retrieved.

Finally, a most important new feature of our algorithm is the evolution of solvent boundary or protein mask. It is true that the boundary could change in the traditional solvent flattening cycles, but not to the extent that happens in our algorithm, where it goes from a completely random boundary to a very accurate one. That is why we can start from random phases and the traditional refinement requires a set of good phases to start with. All of the above factors conspire to make *ab initio* phasing possible. For the same reasons, our iterative scheme can greatly increase the chance of success of a refinement job.

6. Conclusion

The traditional way of solving the phase problem starts by collecting the experimentally determined phases, which are rarely accurate enough to yield an interpretable electron-density map. Phase improvement using a variety of density-modification methods are generally required. Solvent flattening, histogram matching (Zhang & Main, 1990*a*,*b*) and non-crystallographic averaging (Rossmann, 1995) are the main techniques. In general, it is believed (Taylor, 2010) that density-modification techniques will not turn a bad map into a good one, but they will certainly improve a promising map that shows some interpretable features.

It has gradually been realized (Liu *et al.*, 2012; Millane & Lo, 2013) in very recent years that a general class of iterative projection algorithms (Elser, 2003), which includes the HIO scheme, can considerably increase the radius of convergence over the conventional density-modification algorithms. Those algorithms offer the possibility of protein structure determination starting with only information on the molecular envelope (Hao, 2001) and low-order non-crystallographic symmetry.

It turns out that, as we have demonstrated in this paper, the convergence region of the iterative projection algorithms can be so large that no prior knowledge of the molecular envelope is needed at all, at least in high-solvent-content crystals. Almost any given initial density or phases will iterate towards the correct density or phases, given a large enough number of iterations. With modest NCS, the same thing can happen for low-solvent-content crystals. Thus direct phasing is quite likely for most, if not all, protein crystals.

This work was partially supported by the Texas Center for Superconductivity and the Robert A. Welch Foundation (E-1070). We thank Yu-Hui Dong for useful correspondence and G. Phillips Jr for useful conversations.

References

Acharya, P., Goenrich, M., Hagemeier, C. H., Demmer, U., Vorholt, J. A., Thauer, R. K. & Ermler, U. (2005). *J. Biol. Chem.* **280**, 13712–13719.
Bleicher, L. *et al.* (2008). *BMC Struct. Biol.* **8**, 8.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Chen, C. C., Miao, J., Wang, C. W. & Lee, T. K. (2007). *Phys. Rev. B*, **76**, 064113.
Dronyak, R., Liang, K. S., Stetsko, Y. P., Lee, T. K., Feng, C. K., Tsai, J. S. & Chen, F. R. (2009). *Appl. Phys. Lett.* **95**, 11908.
Elser, V. (2003). *Acta Cryst.* A**59**, 201–209.
Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.
Hao, Q. (2001). *Acta Cryst.* D**57**, 1410–1414.
Koepke, J., Krammer, E. M., Klingen, A. R., Sebban, P., Ullmann, G. M. & Fritzsch, G. (2007). *J. Mol. Biol.* **371**, 396–409.
Liu, Z. C. (2012). PhD thesis, Institute of High Energy Physics, Chinese Academy of Sciences, People's Republic of China.
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.
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(R).
Meining, W., Mortl, S., Fischer, M., Cushman, M., Bacher, A. & Ladenstein, R. (2000). *J. Mol. Biol.* **299**, 181–197.
Miao, J., Kirz, J. & Sayre, D. (2000). *Acta Cryst.* D**56**, 1312–1315.
Miao, J. & Sayre, D. (2000). *Acta Cryst.* A**56**, 596–605.
Miao, J., Sayre, D. & Chapman, H. N. (1998). *J. Opt. Soc. Am. A*, **15**, 1662–1669.
Millane, R. P. & Lo, V. L. (2013). *Acta Cryst.* A**69**, 517–527.
Millane, R. P. & Stroud, W. J. (1997). *J. Opt. Soc. Am. A*, **14**, 568–579.
Plas, J. L. van der & Millane, R. P. (2000). *Proc. SPIE*, **4123**, 249–260.
Rossmann, M. (1995). *Curr. Opin. Struct. Biol.* **5**, 650–655.
Strop, P., Brzustowicz, M. R. & Brunger, A. T. (2007). *Acta Cryst.* D**63**, 188–196.
Su, W.-P. (2008). *Acta Cryst.* A**64**, 625–630.
Taylor, G. L. (2010). *Acta Cryst.* D**66**, 325–338.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
Zhang, K. Y. J. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 41–46.
Zhang, K. Y. J. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 377–381.
Zou, G. & Phillips, G. N. Jr (1998). Technical report (CRPC-TR97743). Center for Research on Parallel Computation, Rice University, USA.