Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 28 April 2008
Accepted 27 August 2008

Retrieving low- and medium-resolution structural features of macromolecules directly from the diffraction intensities – a real-space approach to the X-ray phase problem

Wu-Pei Su

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

A simple mathematical algorithm is proposed to generate electron-density functions whose Fourier amplitudes match the diffraction intensities. The function is by construction everywhere positive. Using appropriate averaging procedures, the high-density regions of such functions could yield important structural information about macromolecular crystals. Trial calculations on protein crystals show that the protein envelope plus other structural motifs such as barrels and secondary structures could be recognized in the density maps. As such, the algorithm could provide a basis for new phasing methods or supplement existing phasing methods.

1. Introduction

Despite the availability of experimental methods to solve the phase problem in macromolecular X-ray crystallography, it is still desirable (Hao, 2006; Strop *et al.*, 2007) to translate the diffraction data directly into structural information.

There have been attempts (Subbiah, 1991; David & Subbiah, 1994) at such *ab initio* calculations. One approach by Lunina *et al.* (2003) focused on generating good phases based on connectivity analysis. They showed that for small unit cells, molecular envelopes as well as secondary-structure elements can be identified in the electron-density maps calculated. In another approach (Webster & Hilgenfeld, 2001), more emphasis is placed on the real-space aspect, *i.e.* the density function itself. A regular three-dimensional grid is chosen to occupy the unit cell, upon which a number of uniform scatterers are initially populated randomly. The configuration is then subjected to a series of improvement cycles (*via* an elaborate genetic algorithm) such that the calculated structure amplitudes best match the experimental ones. After that, certain density modification is applied to the density map to arrive at a final map. For the two trial calculations reported, the resultant maps actually describe the solvent rather than the protein regions.

We have followed Webster and Hilgenfeld in adopting a regular grid and appropriately positioned uniform scatterers, but we have employed a very simple method (steepest descent) to improve the match between the calculated and observed Fourier amplitudes. Using a proper averaging procedure, we find that the density function generated so simply contains much structural information including macromolecular envelopes, major motifs such as barrels and even secondary-structure elements. Our calculations were carried out using real diffraction data, thus the finding is of great practical interest.

In the following, a few trial calculations will be presented. After that, we will discuss further extensions and how to integrate this method with other existing phasing methodologies.

2. Methodology

As mentioned in §1, we have followed Webster & Hilgenfeld (2001) in filling the unit cell with a regular grid. Each grid point can have one or zero point scatterers. An electron-density function is thus characterized by the sites with one scatterer. This will be referred to as a configuration. For each configuration the structure factor $F_{\mathbf{k}}$ can be calculated, the magnitude of which, F_1, is to be compared with the observed one, $|F_{\mathbf{k}}|_{\rm obs} = F_2$, up to an overall scale factor λ. The discrepancy is measured by the residual

$$R = \sum_{\mathbf{k}} (F_2 - \lambda F_1)^2. \qquad (1)$$

In equation (1), R is a quadratic function of λ. It can be minimized with respect to λ. Substituting this optimal value of λ into the same equation yields the following expression for R:

$$R = \left[\sum_{\mathbf{k}} F_2^2 \sum_{\mathbf{k}} F_1^2 - \left(\sum_{\mathbf{k}} F_1 F_2 \right)^2 \right] \Big/ \sum_{\mathbf{k}} F_1^2. \qquad (2)$$

Starting from an initial configuration of scatterers, a grid point is picked randomly. The number of scatterers at this grid point is changed if the resultant residual R is reduced. This is followed by another randomly picked grid point. This procedure is repeated until the residual falls below a certain target value. It should be noted that the crystal symmetry should be

respected during the entire procedure, *i.e.* each configuration should be consistent with the symmetry of the unit cell.

In order to exhibit interesting features, the density function obtained after the minimization of the residual needs to be averaged and truncated, similar to what is done in solvent flattening (Woolfson & Fan, 1995). For each grid point occupied by a scatterer, a weighted average of the electron density is calculated by summing over the contributions of all scatterers within a certain distance. The weighting function is the reciprocal distance between the two scatterers. In this fashion, each scatterer carries a certain average density; those scatterers whose average densities are below a cutoff are dropped. The remaining scatterers are shown as small spheres in the figures below.

In the following, we report on three trial calculations. All figures give stereoscopic views.

3. Examples

(1) *Retinol-binding protein (RBP) (PDB ID: 1AQB)*. We have picked this structure partly because the space group $P2_12_12_1$ is a common one. In addition, the unit-cell dimensions are modest: a = 45.81, b = 53.137 and c = 72.966 Å. Real diffraction data are used for the trial calculation. There are 495 reflections below 6 Å resolution and 198 reflections below 8 Å resolution.

First we describe the 8 Å calculation. The grid is $26 \times 26 \times 36$, *i.e.* there are 26 divisions in the a direction *etc*. Initially, about half of the grid points are occupied by a scatterer. This ratio stays pretty constant through the minimization steps. As mentioned before, each step consists of picking a grid point randomly and updating the number of scatterers. After 4000 steps, the residual R drops from 11 to 0.4. In principle, the residual can be arbitrarily lowered given enough iterations. Empirically, we find that too many iteration steps actually degrades the quality of the density map, probably because of overfitting the diffraction data. The optimal number of iterations determined empirically does not seem to vary much from structure to structure.

The configuration obtained at the end of residual minimization is subjected to an averaging procedure with a weighting function which cuts off at about 10 or 12 Å, *i.e.* a scatterer further away from another scatterer does not contribute to the average density of that scatterer. Within that cutoff, the weight is inversely proportional to the distance. With the help of the average density, about half of the scatterers are discarded. The remaining ones are shown in Fig. 1 as green balls. The native structure is shown by the wire frame. The entire figure is a sliced unit cell. It is clear that the green balls correlate well with the native structure in terms of protein envelope and solvent region. There is a complete copy of RBP near the upper-left corner of the unit cell. Fig. 2 gives another view of the same unit cell.

It should be emphasized that the entire calculation takes only a few minutes on a PC, but not every calculation yields a map which agrees so well with the native structure. Nonetheless, the chance of getting a good map is of the order of a few tenths. When faced with an unknown structure, the challenge is to be able to recognize the good maps among the bad

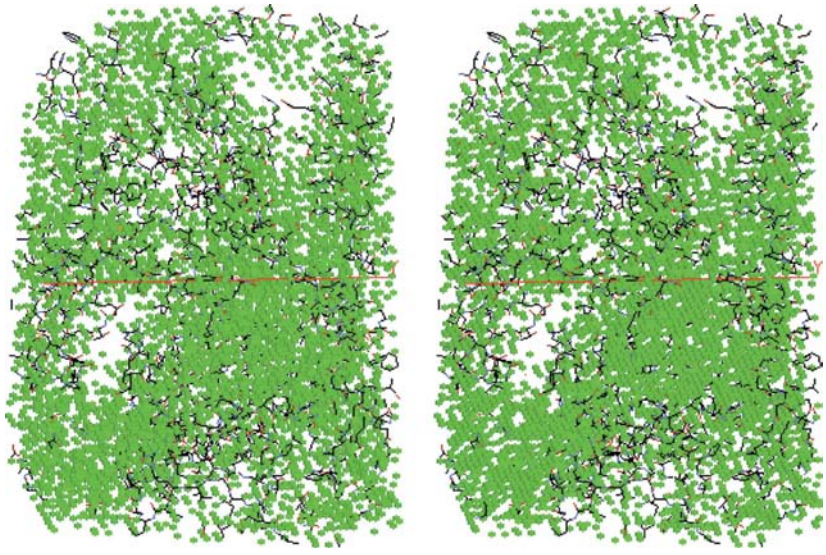

Figure 1
Comparison of a calculated (8 Å) map of RBP with the native structure as a wire frame. Note a complete copy of RBP near the upper-left corner of the unit cell. The y axis points to the right and the z axis points up.

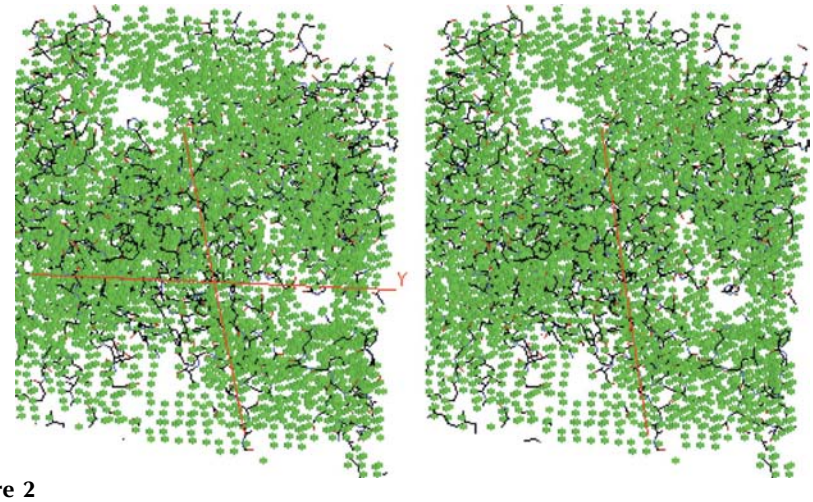

Figure 2
Another view of the maps in Fig. 1.

ones. We will discuss more of this later. For this structure, let us mention that we have originally tried to solve it without prior structural knowledge and indeed were able to identify a copy of it near the upper-left corner of the unit cell with a proper choice of origin and enantiomorph.

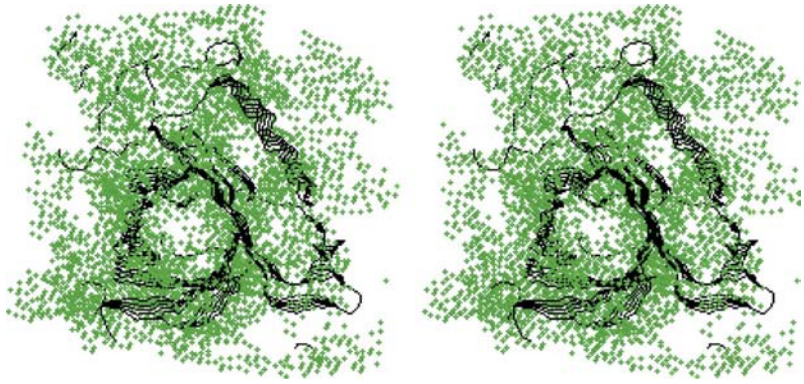

Figure 3
A calculated (6 Å) map of RBP compared with the native structure as a strand diagram. Only a portion of the unit cell is shown.

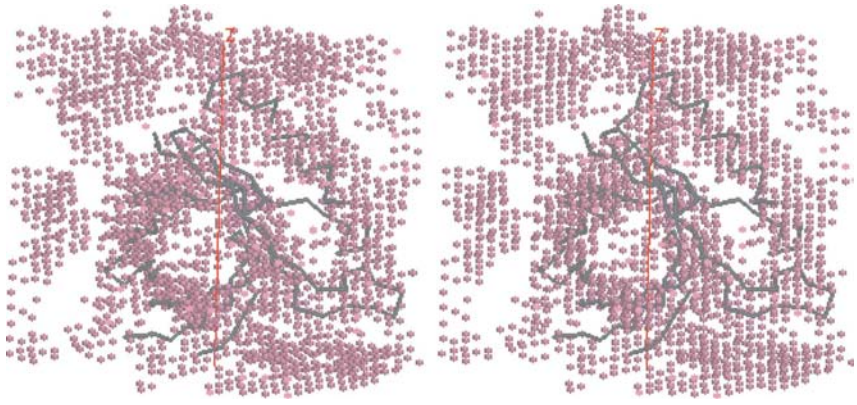

Figure 4
Another calculated (6 Å) map of RBP revealing an α-helix flanking the barrel.

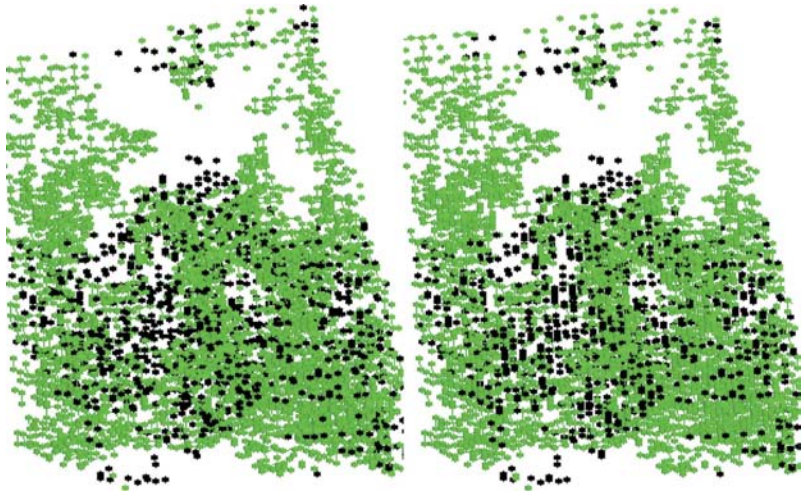

Figure 5
Calculated electron-density maps of EP8. The green map is derived from 5 Å diffraction data, whereas the black map is derived from 8 Å data. Notice the thin vertical slab near the center of the maps.

To reveal more detailed structure, another calculation using 6 Å diffraction data was carried out. To be consistent with a higher resolution, a finer grid, $30 \times 30 \times 48$, is adopted and the weighting function is cut off at a distance of 6 Å. The same number of updates as before is applied to the residual, which drops from 18.7 to 4.4. A portion of the averaged and truncated map is shown in Fig. 3. Again the green dots are from the calculation superimposed on the native structure as a strand diagram. Clearly the green dots outline very well the barrel motif of RBP. Our experience is that this type of structural motif is easily recognizable in an unknown macromolecule.

RBP is dominated by the barrel motif, but it also contains a sizable α-helix flanking the barrel. It shows up clearly in another calculated map shown in Fig. 4. An α-helix has the characteristic cylindrical shape and can be readily recognized in a map. To avoid spurious identification, one looks for a cylinder appearing in exactly the same location in another map. If that happens in many independently generated maps, it is quite likely to be a genuine α-helix. The same principle can be applied to the identification of β-sheets, as will be seen in the next example.

(2) *Mutated thioredoxin (EP8) (PDB ID: 1EP8)*. We have chosen this structure to test the capability of the method in determining the location of a β-sheet. Thus we knew beforehand that the protein has an α/β motif, *i.e.* an open twisted β-sheet surrounded by α-helices on both sides. We also knew that there are two copies of thioredoxin forming a dimer in the asymmetric unit. Other than those pieces of general information, everything else was unknown and had to be calculated. The space group $P3_121$ is special in that there are only two choices of the origin. The unit-cell dimensions are $a = b = 49.692$ Å and $c = 145.551$ Å.

Following a procedure similar to that in example (1), we first identified a copy of thioredoxin in a particular corner of the unit cell. We then focused on that region and examined each density map for possible signs of a β-sheet. Using diffraction data up to 5 Å, with a $28 \times 28 \times 84$ grid and 5000

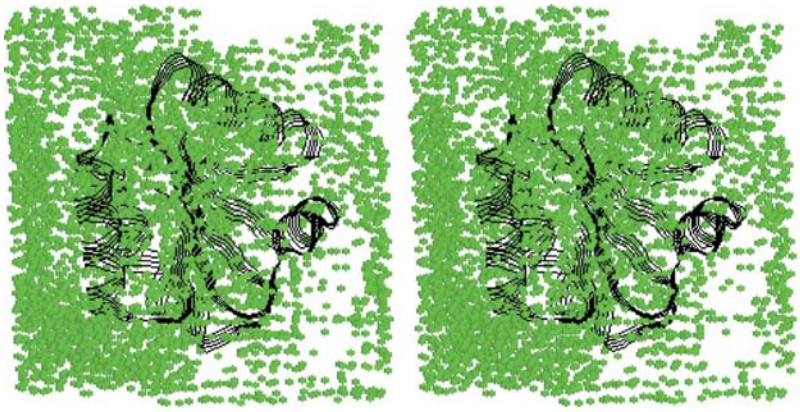

Figure 6
A calculated electron-density map of EP8 compared with a strand diagram of the native structure, viewed at an angle different from that of Fig. 5.

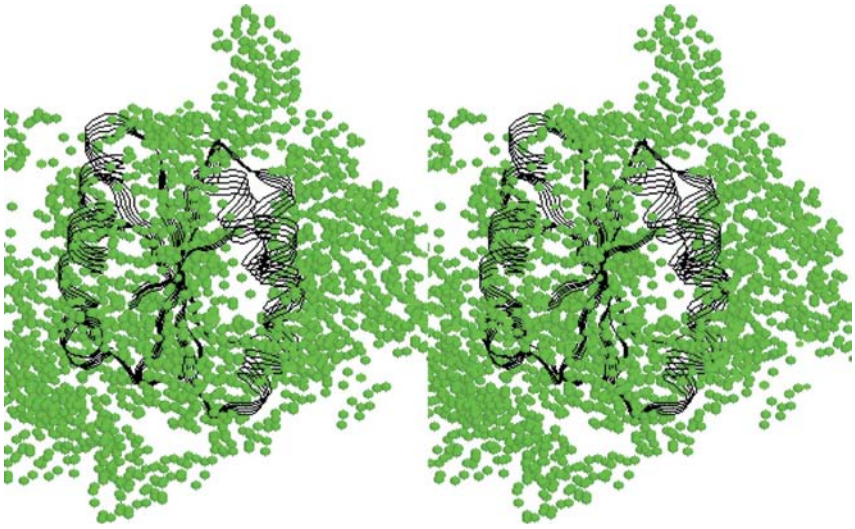

Figure 7
Comparison of another calculated β-sheet profile with the native structure of EP8.

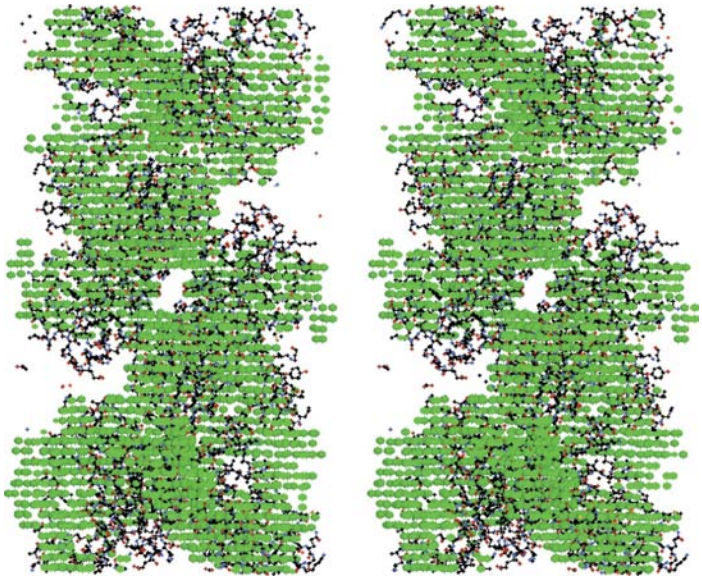

Figure 8
A calculated (7 Å) map of EP8 superimposed on a wire-frame diagram of the native structure.

updates, we were able to generate maps with recognizable features of a β-sheet. An example is shown in Fig. 5. The green dots clearly suggest a thin vertical sheet near the middle of the diagram, which shows only a portion of the unit cell. The black dots are derived from 8 Å diffraction data; they were helpful in locating a copy of thioredoxin. They provide a consistency check whether a feature found with 5 Å data also survives with lower-resolution 8 Å data. After repeatedly seeing the same large flat high-density region in the same location of the unit cell, we were sure that there must be a β-sheet there. A superimposition of the calculated map with the strand diagram of the native structure is displayed in Fig. 6. Indeed there is a very good agreement in terms of the β-sheet profile. Another example of the good agreement is shown in Fig. 7. This is a clear demonstration of the potential of the current method in locating a β-sheet in a new structure.

Although we succeeded in locating the β-sheet of one copy of thioredoxin in our original calculation, we did not correctly locate the second copy then. Later a map generated with 7 Å diffraction data was found to agree moderately well with the native structure in terms of protein envelope, as shown in Fig. 8. Unlike the secondary structures, which possess definite characteristics and can therefore be unambiguously identified in a calculated map, the precise envelope of the protein is harder to pinpoint. Thus it would be hard to tell how good the calculated map in Fig. 8 is without making a comparison with the native structure. It probably would be more definitive to look for a second β-sheet in the unit cell and use that to locate the second copy of thioredoxin. Other criteria include resolution consistency, *i.e.* whether the same envelope emerges in a higher-resolution calculation, and other general constraints such as contacts between different copies.

In retrospect, we could also have used the β-sheet as a pivot or reference for elucidating the α-helices next to the β-sheet. It is easier to spot an

α-helix once we know where to look for it. In this case, right next to the β-sheet. In addition, we could also superimpose two independently calculated maps as shown by different colors in Fig. 9. The two maps have substantial overlap and they also complement each other to yield a more complete picture of the α/β motif.

(3) *Diphtheria toxin repressor (DTXR) (PDB ID: 1G3W)*. As our last example, we choose a rather difficult structure. It is difficult to image partly because of the long chain, the irregular shape of the protein and the large solvent content. In addition, the quality of the diffraction data is probably not as good as those of the previous examples. The space group is the same as the second example, $P3_121$. The unit-cell dimensions are $a = b = 63.5$ Å and $c = 110.0$ Å.

Among its complicated components, DTXR contains a few long α-helices. Our purpose in considering this structure is to show that despite the difficulties mentioned above, maps can be generated from the diffraction data which reveal the long α-helices. An example is shown in Fig. 10. The green map is calculated with 6 Å diffraction data. The grid is $34 \times 34 \times 60$. The number of updates is 5000. The longest α-helix overlaps very well with the calculated map, whereas the rest of the structure does not match the calculation. Although such an extended cylindrical feature in a calculated map is unlikely to be spurious, it is important to confirm it by observing it in the same location inside the unit cell in another independent calculation. Fig. 11 gives such a confirmation. It also reveals another helix intersecting with the longest one. It is actually the second longest one. This latter helix is again revealed more decisively in another calculated map, shown in Fig. 12.

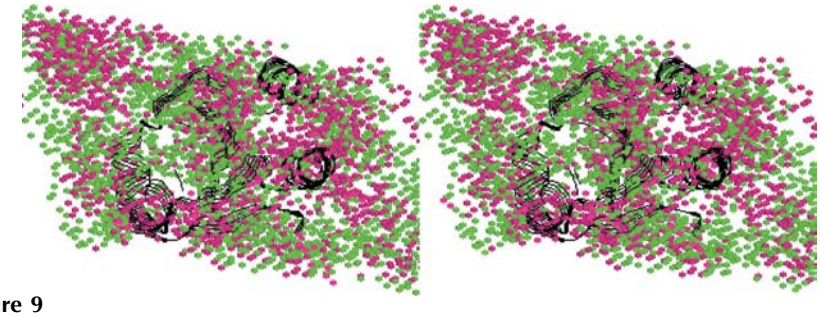

Figure 9
Two independently calculated maps of EP8 superimposed on the native structure as a strand diagram.

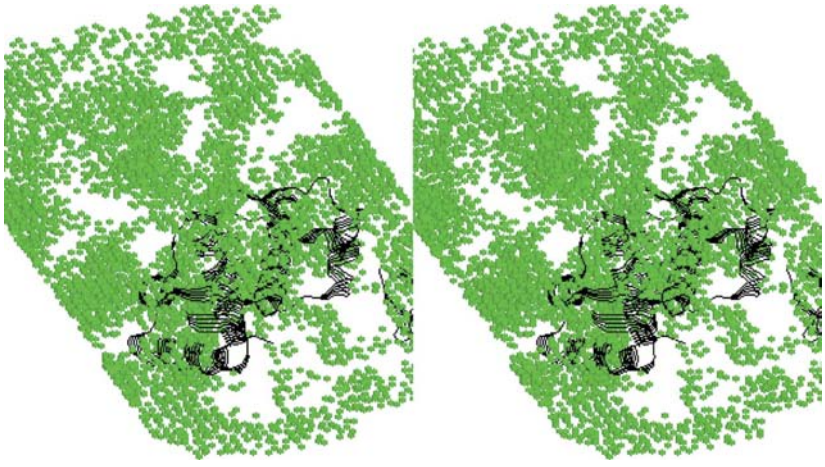

Figure 10
A calculated (6 Å) map of DTXR showing the major α-helix.

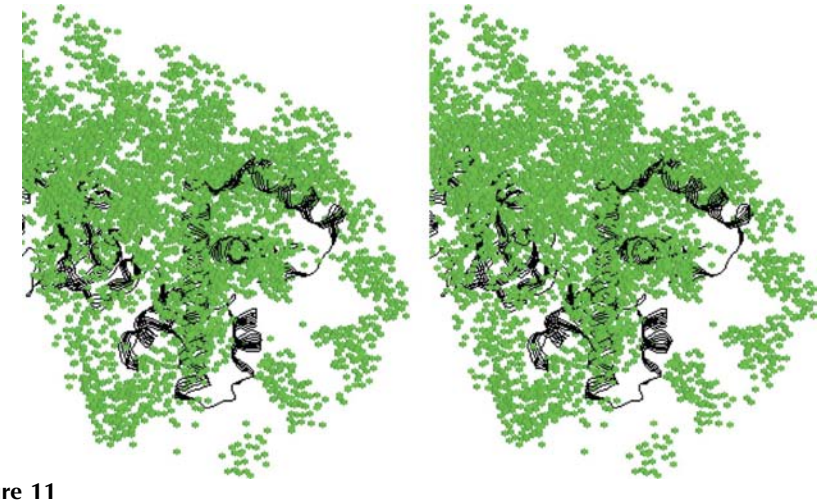

Figure 11
A calculated map of DTXR revealing two intersecting α-helices.

A few comments are in order. First, just like in the previous examples, not all calculated maps show long cylinders as in Figs. 10–12. As a matter of fact, only less than 10% of the generated maps do so. Secondly, as emphasized before, since an α-helix is a well defined object, it can be nailed down given enough good maps. Thirdly, the helices determined from different calculations can be assembled together to form a larger partially solved structure. In this case, the intersecting longest and second longest α-helices are confirmed by three maps. That constitutes a method for extending a small partial structure into a larger one.

4. Discussion

Through the examples given above, it is clear that we indeed have a simple algorithm to generate (directly from the intensity data) density maps that

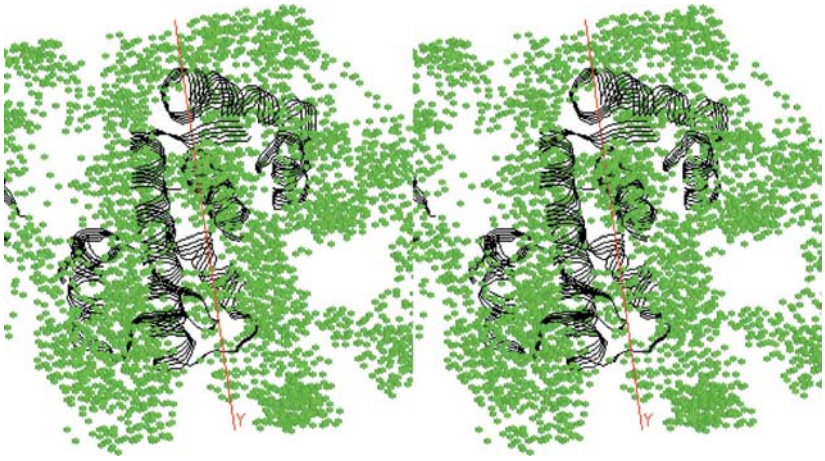

Figure 12
A calculated map of DTXR showing the second longest α-helix near the y axis.

contain useful structural information on macromolecules. The algorithm also generates 'bad' maps which could confuse the user when looking for good maps. For the algorithm to be of common usage in solving new structures, better screening tools (preferably automatic) need to be developed to recognize good maps. So far we have done the screening manually. We have had varying degrees of success for trial calculations involving structures without prior information. There are universal features associated with secondary structures, therefore they are easier to spot. For molecular envelopes and other less common motifs, more intuition or extra constraints are required to narrow the set of criteria. Small-angle X-ray solution scattering can provide a source of information on the envelope (Hao *et al.*, 1999; Hao, 2001), for example.

In many cases, even a very good map usually contains only partially correct structures. Thus one might be able to determine one helix from one set of maps and another helix from another set of maps *etc.* An advantage of a real-space approach is that one can easily combine smaller partial structures into a larger partial structure, as we have seen in example (3). In this fashion, one might be able to construct a major portion of the final complete structure.

Aside from straightforward extension in real space, phase extension offers another useful alternative. In the conventional direct method (Karle, 1968) a partial atomic structure generates a good set of starting phases, which can be further extended through the tangent formula. A new Fourier synthesis typically yields a larger atomic structure, *i.e.* with more atoms located. The tangent formula is not useful for macromolecules, but there are other tools for phase extension through noncrystallographic symmetry averaging and density modifications including solvent flattening and histogram matching (Woolfson & Fan, 1995). A good example of phase extension starting from core α-helices of a membrane protein was given by Strop *et al.* (2007).

Another useful idea in this connection is the mask strategy (Chacón *et al.*, 1998). An image obtained at a lower resolution can be used as a mask to guide (Su, 1995; Chou & Lee, 2002) a more refined search at a higher resolution. Thus instead of starting a 6 Å calculation from scratch, it may be easier to input an 8 Å image to limit the possible configurations. Those configurations that deviate too much from the 8 Å image are not considered in the minimization of residual. The connectivity criterion emphasized by Lunina *et al.* (2003) is another constraint to impose.

Whatever future extensions might be, we have achieved an important step in extracting very useful structural information directly from the diffraction intensities. Much remains to be explored.

This work was partially supported by the Texas Center for Superconductivity and the Robert A. Welch Foundation (E-1070). I thank Quan Hao for useful conversations.

References

Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andrew, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
Chou, C. I. & Lee, T. K. (2002). *Acta Cryst.* A**58**, 42–46.
David, P. R. & Subbiah, S. (1994). *Acta Cryst.* D**50**, 132–138.
Hao, Q. (2001). *Acta Cryst.* D**57**, 1410–1414.
Hao, Q. (2006). *Acta Cryst.* D**62**, 909–914.
Hao, Q., Dodd, F. E., Grossmann, J. G. & Hasnain, S. S. (1999). *Acta Cryst.* D**55**, 243–246.
Karle, J. (1968). *Acta Cryst.* B**24**, 182–186.
Lunina, N., Lunin, V. & Urzhumtsev, A. (2003). *Acta Cryst.* D**59**, 1702–1715.
Strop, P., Brzustowicz, M. R. & Brunger, A. T. (2007). *Acta Cryst.* D**63**, 188–196.
Su, W.-P. (1995). *Physica (Utrecht) A*, **221**, 193–201.
Subbiah, S. (1991). *Science*, **252**, 128–133.
Webster, G. & Hilgenfeld, R. (2001). *Acta Cryst.* A**57**, 351–358.
Woolfson, M. & Fan, H.-F. (1995). *Physical and Non-physical Methods of Solving Crystal Structures*. Cambridge University Press.