

Contributed Papers

WHY PROTEINS CAN BE SOLVED BY DIRECT METHODS

M.M. WOOLFSON

Department of Physics, University of York, York YO1 5dd, U.K.

1. Recent Work

Woolfson and Yao [1] were the first to show that it was possible to solve a small protein by the use of direct methods. The protein they used as a demonstration, avian pancreatic polypeptide (aPP) contained 301 atoms in the asymmetric unit, including one zinc atom, and so was less than twice as large as small-molecule structures which had previously been solved. However, Yao and Woolfson only knew that good phase sets had been obtained by the direct-method program SAYTAN because the structure had previously been solved. At that time no effective figures-of-merit were available to distinguish the better sets of phases. Other successful demonstrations of solving protein structures followed-e.g. Sheldrick et al. [21, Hauptman [3], Mukherjee & Woolfson [4], 5], Mukherjee [6] and Mukherjee, Ghosh & Woolfson [7].

2. Theoretical Considerations

The various applications all depended in some way on the application of the Cochran phase relationship [8].

$$\Phi_3(h, k) = \phi(h) - \phi(k) - \phi(h - k) \quad (1)$$

$$\approx 0 \pmod{2\pi}$$

Where \approx means 'is distributed around'. The form of the Cochran distribution is shown in Fig. 1a. The equation for it is

$$P(\Phi_3) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos \Phi_3). \quad (2)$$

The standard deviation of the Cochran distribution depends on

$$\kappa(h, k) = 2N^{1/2} |E(h)E(k)E(h - k)| \quad (3)$$

where there are N equal atoms in the unit cell and the E 's are normalized structure factors.

The maximum value of the product of three E 's for any structure is usually about 20 so for a structure with, say, 100 atoms in the unit cell the maximum

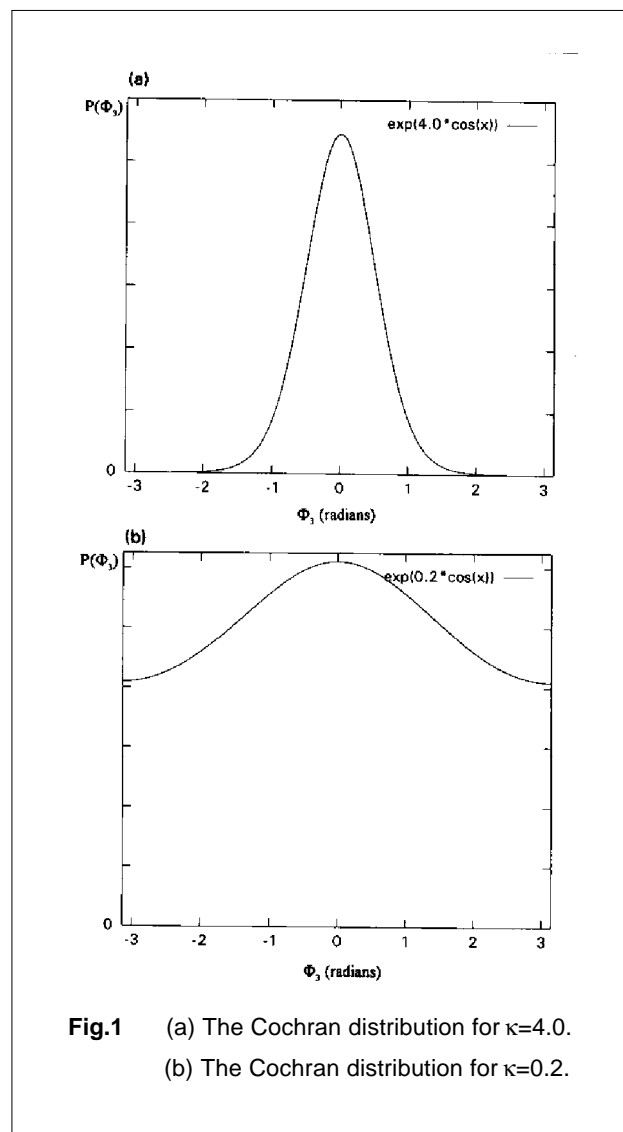


Fig.1 (a) The Cochran distribution for $\kappa=4.0$.
 (b) The Cochran distribution for $\kappa=0.2$.

value of κ is about 4, corresponding to a standard deviation of about 30° . On the other hand for a small protein with 3,600 atoms in the unit cell the maximum value of κ falls to about 0.6 with a corresponding standard deviation of about 80° . Many of the three-phase relationships which have to be used by a direct method in solving a protein, even a small one, would have κ values about 0.2 or so; the Cochran distribution for $\kappa=0.2$ is shown in Fig. 1b and is seen to be fairly flat.

In terms of the validity of individual phase relationships the prospects for solving protein structures look quite bleak although, to compensate for the weakness of the individual relationships, there are many more relationships available to be used. However, there is another way of looking at the application of direct methods. Cochran's approach to finding the form of the three phase relationship (originally the sign relationship [9]) was to consider the expected appearance of an electron-density map. Maps calculated with any sets of phases, even random phases, would all have the same average and standard deviation of the density distribution. For an E-map, calculated with E's as coefficients these are given by

$$\bar{\rho} = \frac{E(0)}{V} \quad (4a)$$

and

$$\sigma_p = \frac{1}{V} \left\{ \sum_{h \neq 0} |E(h)|^2 \right\}^{1/2}. \quad (4b)$$

Cochran expressed the condition for a good phase set as being that the corresponding electron-density map should show density concentrated around atomic centres with flat non-zero density inbetween which he interpreted as that of making $\int_V \rho^3 dV$ a maximum. It is easily shown that

$$\int_V \rho^3 dV = \frac{1}{V^2} \sum_h \sum_k |E(h)E(k)E(h-k)| \times \cos\{\phi(h) - \phi(k) - \phi(h-k)\} \quad (5)$$

and the three-phase relationship comes from maximizing the cosine terms on the right-hand side. For small structures the summation in (5) is close to its global maximum and maximizing the summation is efficiently carried out by means of the tangent formula [10]. However, for proteins the summation for correct phases is very far from its global maximum so this raises the question of how it is that a direct method based on the application of the tangent formula can lead to the structure.

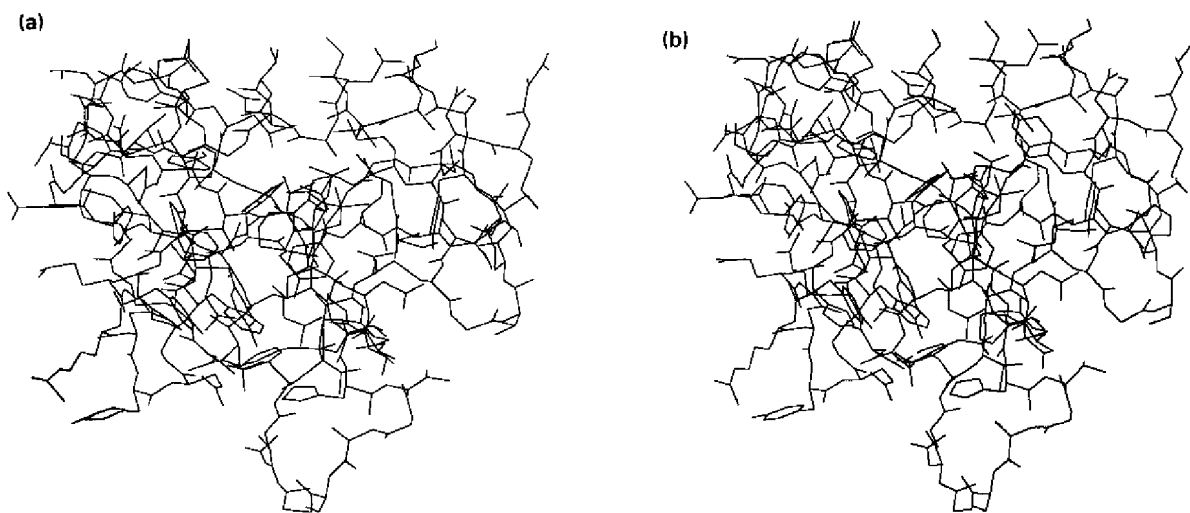


Fig. 2. (A) The structure of RNAP1 from a direct method. (b) The published structure of RNAP1.

3. Finding and Using Fragments

Many of the sets of phases found by direct methods for proteins have phases which satisfy the phase relationships far better than the true phases and sets which are closest to correct phases may have a mean phase error of about 70° . Indeed this was the mean phase error found for the previously-known structure RNAP1 by Mukherjee, Ghosh and Woolfson [8] and the final structure found, compared with the structure from published coordinates, is shown in Fig. 2. To understand how it is that the structure can be found from such an unpromising beginning, in terms of mean phase error, it is best to think in terms of Cochran's condition on the electron density. If the Cochran integral in (5) is large then one way this can be so is if the density is concentrated around a number of centres and is small elsewhere. There are, of course, other ways for the Cochran integral to be large—for example, by creation of one, or a few, very large peaks even though density may be negative in other regions. In practice there is a tendency for an appearance of atomicity in the maps found for proteins by the application of direct methods and we now consider the implication of this.

The Patterson function is the convolution of the electron density with its own inverse or, in terms of structure, a map showing the interatomic vectors. Every density map, regardless of the phases used to create it, has the same corresponding Patterson function since the Patterson function is phase independent. Hence if the map has the quality of atomicity then a large proportion of the vectors between the peaks in the map must correspond to true peaks in the structure. The most effective way for interpeak vectors in the map to correspond with true vectors in the structure is if the map contains images of fragments of the structure in the correct orientation, albeit not in the correct position. Refaat and Woolfson [11] used this concept in a procedure they called TRITAN. This was a way of recycling information from unsuccessful runs of the direct-method programme MULTAN. Maps were calculated for a few of the phase sets with better figures of merit - which were likely to have the greatest atomicity. The largest fragment was taken from each map and used to give estimates of three phase invariants. The estimates from different fragments were combined to give overall estimates which were then inserted into a new run of MULTAN.

In all the cases tried this yielded the structure even although, for one of the trial structures, the phase sets with the worst figures of merit were used.

4. The Way Ahead

In all cases the largest structures solved so far by direct methods have depended on gradually building up the structure from some of the peaks found in original maps. It seems very unlikely that direct methods will ever be developed to the extent that the first map will show the majority of the peaks for a protein with 1,000 or more independent non-hydrogen atoms; it is much more likely that some fragment will be recognised and used as a starting point.

It thus seems that the problem of solving protein structures by direct methods can be divided into two parts. The first is to produce the best possible map for getting a fragment. Present methods tend to give phases which satisfy the phase relationships too well. Is this a good thing, in that it maximizes the Cochran integral and hence might give atomicity, or would it be better to generate phases which satisfy the relationships to the theoretical extent but not more than that? This would make the Cochran integral smaller and hence give less tendency for atomicity, but could there be compensation for this by including other conditions which independently militate against negativity? Again, would phase sets be improved by developing phases to satisfy the Sayre equation [12]? The first small protein solution was by SAYTAN, which actively uses the Sayre equation, but the much larger structure of RNAP1 just used a modified MULTAN approach.

If the best, presumably largest, fragment is found then arises the problem of developing this to the complete structure. There are several different procedures for this which have been used successfully but for larger proteins it may be difficult to find a fragment large enough to act as a starting point. In this case trying to amalgamate information from several phase sets may be an advantage.

The TRITAN method, which uses the phases from the fragments in a reciprocal-space approach, is clearly one possible way. Another possibility is to try to combine fragments in real space to provide a single larger fragment. This has the difficulty that fragments from different phase sets may have very little or no

overlap but if a heavy atom is present and can be detected in a number of fragments then this can be used as an anchor point for the combination.

The development of direct methods towards the solution of protein structures is certainly an interesting and challenging exercise for those who are involved but it is unlikely to make a major impact of protein crystallography. Physically-based techniques of solving protein structures have served the community of protein crystallographers quite well - to the point where virus structures are being investigated. The extension of anomalous scattering to seleno-proteins, where selenium has replaced sulphur, or even anomalous scattering from sulphur itself, may leave very few proteins for which direct methods would be the only method to use. A few cases may arise where direct methods may fill a gap but these will be few. The most promising future for direct methods in the protein field is if they develop to the point where for, say, 2,000 atom problems they give automatic and quick solutions with powerful computers and they would then become the method of choice.

REFERENCES

- [1] M. M. Woolfson & Yao Jia-xing. (1990) *Acta Cryst.* **A46**, 11-46.
- [2] G. M. Sheldrick, Z. Dauter, K. S. Wilson, H. Hope, & L.C. Seiker (1993) *Acta Cryst.* **D46**, 18-23.
- [3] H. Hauptman (1995) *Acta Cryst.* **B51**, 416-422.
- [4] M. Mukherjee, & M. M. Woolfson (1993) *Acta Cryst.* **D49**, 9-12.
- [5] M. Mukherjee & M. M. Woolfson (1995) *Acta Cryst.* **D51**, 626-628,
- [6] M. Mukherjee (1998) Submitted to *Acta Cryst. D*.
- [7] M. Mukherjee, S. Ghosh & M. M. Woolfson (1998) Submitted to *Acta Cryst. D*.
- [8] W. Cochran (1955) *Acta Cryst.* **8**, 473-478.
- [9] W. Cochran (1952) *Acta Cryst.* **5**, 65-67.
- [10] J. Karle & H. Hauptman (1956) *Acta Cryst.* **9**, 635-651.
- [11] L. S. Refaat & M. M. Woolfson (1988) *Acta Cryst.* **A44**, 349-353.
- [12] D. Sayre (1952) *Acta Cryst.* **5**, 6-65.