Acta Cryst. (1993). D49, 13–17

On the Application of Phase Relationships to Complex Structures. XXXIII. The Problems with Large Structures and Low Resolution

BY M. M. WOOLFSON

Department of Physics, University of York, York YO1 5DD, England

(*Received* 11 *May* 1992; *accepted* 11 *September* 1992)

Abstract

A conventional direct method, using the Sayre equation as a basis, has been shown to be capable of solving a small protein with data of 3.0 Å resolution or better. An analysis of the Sayre equation, with data of various resolutions and with different lower limits of $|E|$ for the contributors in the summation, shows that its effectiveness for phasing is independent of structural complexity but does decline as the resolution becomes worse. It is suggested that a practicable lower limit for the application of conventional direct methods is about 3.5 Å. For large macromolecular structures the number of contributors to the summation in the Sayre equation becomes too large to handle and it is suggested that real-space methods should be used instead.

Introduction

Direct methods, in the form of easily used computer packages, have been available for more than 20 years and the great majority of small structures can be readily solved by their use. While there are still occasional difficult small structures which exercise the skill of the crystallographer, it is a valid generalization to state that the standard small-structure problem in crystallography has been virtually solved. Even if direct methods do not provide a solution then there are Patterson-search methods which are also very effective.

The new challenges which now arise in crystallography are in the macromolecular area. This is a field which has been enjoying great success for more than three decades advanced by the ingenious exploitation of physical methods such as isomorphous replacement and, increasingly, anomalous scattering. However, the addition of heavy-atom-containing residues to a protein only gives isomorphism to low resolution and heavy-atom derivatives may not be available for some small proteins. Again, many proteins do not contain convenient anomalous scatterers and cannot accommodate them – although it must be said that current experiments with long wavelengths from synchrotron sources offer the possibility of useful anomalous scattering from sulfur, a common protein ingredient. Here we shall be taking a look at the problems associated with the application of direct methods to proteins and, by understanding the nature of the problems, suggest the best way forward for future work.

A basic analysis of the use of phase relationships

The basis of most direct methods is the tangent formula (Karle & Hauptman, 1956) which may be written in the form

$$\varphi(\mathbf{h}) = \text{phase of } [\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k})]. \qquad (1)$$

However, the method *SAYTAN*, which has been shown to be applicable to a small protein with high-resolution data (Woolfson & Yao, 1990), is based instead on Sayre's equation (Sayre, 1952)

$$F(\mathbf{h}) = [f(\mathbf{h})/g(\mathbf{h})V]\sum_{\mathbf{k}} F(\mathbf{k})F(\mathbf{h}-\mathbf{k}) \qquad (2)$$

where $f(\mathbf{h})$ and $g(\mathbf{h})$ are the scattering factors for normal and squared electron density and V the volume of the unit cell. The equation is only strictly valid for equal resolved atoms although Shiono & Woolfson (1991) have shown that it also applies quite well over a range of conditions, including for structures at low resolution and with unequal atoms. It will be seen that the tangent formula is equivalent to using just the phase information from the right-hand side of Sayre's equation.

The equation can be used to link normalized structure factors (the E's) which are normally employed in direct methods since, with enough data, an E map shows well resolved atoms despite diffraction ripples. In this case, for an equal-atom structure,

$$f(\mathbf{h}) = N^{-1/2} \qquad (3)$$

and $g(\mathbf{h})$, which is the self convolution of the f's in reciprocal space, is given by a summation

$$g(\mathbf{h}) = V^*\sum_{\mathbf{k}} f(\mathbf{k})f(\mathbf{h}-\mathbf{k}) \qquad (4)$$

where V^* is the volume of the reciprocal lattice cell. Since $VV^* = 1$ then, inserting the results (3) and (4) in (2), we have Sayre's equation for normalized

0907-4449/93/010013-05$06.00

Table 1. *The range of values of Q for different resolutions for* $|E(\mathbf{h})| = 1.4$ *and structures with an average number of non-hydrogen atoms per unit volume in the crystal*

Q_{max} corresponds to a reflection very close to the origin of reciprocal space while Q_{min} corresponds to a reflection at the resolution limit.

	Resolution (Å)							
	0.77	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Q_{max}	31.5	20.7	10.8	6.7	4.3	2.96	2.05	1.44
Q_{min}	16.7	11.4	5.6	3.3	1.9	1.13	0.61	0.26

structure factors

$$E(\mathbf{h}) = [N^{1/2}/M(\mathbf{h})]\sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k}) \qquad (5)$$

where $M(\mathbf{h})$, the total number of terms in the summations in both (3) and (4), depends on M_{tot}, the total number of reciprocal lattice points within the resolution limit, and the position of the point $\mathbf{h}$ in reciprocal space. Equation (4) was first given by Hughes (1957) in the form

$$E(\mathbf{h}) = N^{1/2}\overline{E(\mathbf{k})E(\mathbf{h}-\mathbf{k})}^{\mathbf{k}}. \qquad (6)$$

In *SAYTAN* (5) is used in a quantitative way to find phases for the larger structure factors which will satisfy the equations for both the large $E(\mathbf{h})$ and also a selection of small (ideally zero) E's. It is known that for small structures most three-phase relationships have values clustered around zero (modulo 2π) and this implies that the components in the summation are fairly well lined up in the complex plane. This then is the necessary condition to satisfy Sayre's equation. Equally it is also well known that for very large structures the three-phase invariants have fairly flat distributions between π and $-\pi$ with only a modest hump in the region of zero. The contributors in the complex plane must, therefore, add up in a way which looks fairly random although, obviously, there must be a bias which leads to satisfying (5). We are now going to examine these ideas in a more quantitative way. The approach will be to look at the distribution of magnitudes which would result from using random phases on the right-hand side of (5) to see how likely it is that an equation for an $E(\mathbf{h})$ of average magnitude could be satisfied just by chance. If the correct magnitude could easily be obtained with random phases then this might imply that the set of equations is not very restrictive and, therefore, that its potential for phase determination is low.

We now write

$$X = \sum_{\mathbf{k}} E(\mathbf{k})E(\mathbf{h}-\mathbf{k}) = \sum_{\mathbf{k}} |\eta(\mathbf{h},\mathbf{k})|\exp[i\psi(\mathbf{h},\mathbf{k})] \qquad (7)$$

and consider the probability distribution of $|X|$ with random ψ's. This problem has already been solved by Wilson (1949) since $|X|$ has the distribution of a structure factor for a non-centrosymmetric structure with $M(\mathbf{h})$ atoms with scattering factors $|\eta(\mathbf{h},\mathbf{k})|$ and positions giving the phase factors in (7). From this we find the distribution

$$P(|X|) = (2/\Sigma)|X|\exp[-|X|^2/\Sigma] \qquad (8)$$

where

$$\Sigma = \sum_{\mathbf{k}} |\eta(\mathbf{h},\mathbf{k})|^2 = \sum_{\mathbf{k}} |E(\mathbf{k})|^2 |E(\mathbf{h}-\mathbf{k})|^2.$$

Assuming that the values of $|E(\mathbf{k})|^2$ and $|E(\mathbf{h}-\mathbf{k})|^2$ are uncorrelated then

$$\Sigma = M(\mathbf{h})\overline{|E(\mathbf{k})|^2}\,\overline{|E(\mathbf{h}-\mathbf{k})|^2} = M(\mathbf{h}) \qquad (9)$$

since, by definition, $\overline{|E|^2} = 1$.

For this distribution $\overline{|X|} = [2M(\mathbf{h})/\pi]^{1/2}$ and $\overline{|X|^2} = \Sigma = M(\mathbf{h})$ so that the standard deviation of the distribution is

$$\sigma_x = (\overline{|X|^2} - \overline{|X|}^2)^{1/2} = 0.603M(\mathbf{h})^{1/2}. \qquad (10)$$

As a measure of the likelihood that $|X|$ could attain a value which would satisfy the magnitude component of the Sayre equation we now find how many standard deviations $|X|$ is from the required value. This is

$$Q = \left\{\frac{M(\mathbf{h})|E(\mathbf{h})|}{N^{1/2}} - \left[\frac{2M(\mathbf{h})}{\pi}\right]^{1/2}\right\} \Big/ 0.603M(\mathbf{h})^{1/2}$$

$$= 1.66\left\{\left[\frac{M(\mathbf{h})}{N}\right]^{1/2}|E(\mathbf{h})| - \left(\frac{2}{\pi}\right)^{1/2}\right\}. \qquad (11)$$

Assuming that a direct-methods approach would use all the data and all possible relationships then, for a given resolution, the surprising result is found that the values of Q would not depend on the structural complexity. For a given number density of atoms in the unit cell and data resolution the ratio M_{tot}/N is constant and $M(\mathbf{h})/M_{tot}$ is the ratio of (the common volume of two reciprocal lattice limiting spheres with centres separated by the vector $\mathbf{h}$):(the volume of one of the spheres). For a structure with data to a resolution of 0.77 Å, the Cu $K\alpha$ limit, the ratio M_{tot}/N will be about 200 and $M(\mathbf{h})/N$ will vary between about 60 and 200. The range of values of Q for this resolution is given in the first column of Table 1. It is clear that the likelihood of satisfying the magnitude requirements of Sayre's equation for greater than average $|E(\mathbf{h})|$ with random phases is extremely low.

For the resolutions usually available with protein data, the ratios of $M(\mathbf{h})/N$ will be much smaller, and Table 1 shows the ranges of Q for various resolutions. For some of the low-resolution columns it appears that a random starting set of phases will be fairly close to satisfying the magnitude aspect of Sayre's equations and possibly a small amount of processing with a tangent-formula approach will

soon reach magnitudes close to the true ones while also giving phase consistency. We shall now look further at this question, taking an empirical approach and also the more realistic situation where only a subset of the largest E's are used in the phasing process.

Using subsets of data

Mukherjee & Woolfson (1992) have applied *SAYTAN* at various resolutions to the structure avian pancreatic polypeptide (aPP), the trial structure used by Woolfson & Yao (1990). This structure, with space group $C2$ and $Z = 4$, contains in the asymmetric unit a 36 amino-acid peptide plus Zn plus 80 H_2O. The results obtained are shown in Table 2 and are derived from 1000 trials, starting with random phases, at each resolution. As expected the quality of the result obtained (measured in terms of mean phase error) deteriorates with the resolution, although the best sets of phases at each resolution give meaningful electron-density maps which provide, at least, a useful basis for fitting models. For example, the conventional correlation coefficients for the maps obtained at 1.77 and 2.0 Å resolution are 0.58 and 0.52 respectively. Table 2 shows different patterns of the number of reflections and the number of relationships together with different values of $E_{\min}$ (the least value of $|E|$ for the subset of reflections whose phases are being sought). These patterns were the result of trial-and-error in finding the best conditions at each resolution. It will be shown that a systematic explanation can be given for these apparently haphazard patterns.

We consider Sayre's equation written in a modified form

$$|E(\mathbf{h})| = \frac{N^{1/2}}{M(\mathbf{h})}\sum_{\mathbf{k}}|E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|\cos\Phi_3(\mathbf{h},\mathbf{k}) \quad (12)$$

where

$$\Phi_3(\mathbf{h},\mathbf{k}) = \varphi(\mathbf{h}) - \varphi(\mathbf{k}) - \varphi(\mathbf{h}-\mathbf{k}). \quad (13)$$

We now consider the value of the summation in (12) if only terms for which $|E| \geq |E|_{\min}$ are included. Before we do this we shall derive some results based on the acentric distribution of E's given by Wilson (1949). This is, for E's

$$P(|E|) = 2|E|\exp(-|E|^2). \quad (14)$$

It is easily found that the proportion of E's with $|E| \geq |E|_{\min}$ is $\exp(-|E|^2_{\min})$ and that their average value is

$$\overline{|E|^2}^{|E| \geq |E_{\min}|} = 1 + |E|^2_{\min}. \quad (15)$$

We now write

$$X = \sum_{\mathbf{k},p}|E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|\cos\Phi_3(\mathbf{h},\mathbf{k}) \quad (16)$$

where the p under the summation sign indicates that the E's in the summation all satisfy the condition $|E| \geq |E|_{\min}$. To estimate the effect of the limited number of terms in the summation we shall assume that the reduction factor, r, due to the limit in $|E|$ is the same for the true values of the terms as can be found by considering their expectation values. While this cannot be strictly true it should be a reasonable approximation.

From (16) we may write

$$\overline{X} = \sum_{\mathbf{k},p}|E(\mathbf{k})||E(\mathbf{h}-\mathbf{k})|\overline{\cos\Phi_3(\mathbf{h},\mathbf{k})} \quad (17)$$

where we assume that the individual magnitudes are known and the expectation value of the cosine term is given by

$$\overline{\cos\Phi_3(\mathbf{h},\mathbf{k})} = I_1(\kappa)/I_0(\kappa) \quad (18)$$

where $I_1(\kappa)$ and $I_0(\kappa)$ are modified Bessel functions and

$$\kappa = 2N^{-1/2}|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|. \quad (19)$$

For large structures the values of κ will be small and then the approximation may be used

$$I_1(\kappa)/I_0(\kappa) = \kappa/2. \quad (20)$$

With this approximation, for a particular structure and $E(\mathbf{h})$, we may write that

$$\overline{X} = N^{-1/2}|E(\mathbf{h})|\sum_{\mathbf{k},p}|E(\mathbf{k})|^2|E(\mathbf{h}-\mathbf{k})|^2. \quad (21)$$

If the terms in the summation are not correlated (and they will be only weakly) then from (15)

$$\overline{X} = N^{-1/2}|E(\mathbf{h})|M(1 + |E|^2_{\min})^2 \quad (22)$$

where M is the number of terms in the partial summation. From this we find the ratio we require which is

$$r = M(1 + |E|^2_{\min})^2/M(h). \quad (23)$$

Thus a modified Sayre equation, with a statistical correction for the limit in the magnitudes of the E's used in the summation, is

$$E(\mathbf{h}) = (N^{1/2}/M)[\sum_{\mathbf{k},p}E(\mathbf{k})E(\mathbf{h}-\mathbf{k})]/(1 + |E|^2_{\min})^2. \quad (24)$$

If we now write, corresponding to (7)

$$Y = \sum_{\mathbf{k},p}E(\mathbf{k})E(\mathbf{h}-\mathbf{k}) \quad (25)$$

then, again using Wilson statistics, if random phases are used on the right-hand side of (25) then $|Y|$ has a probability density with

$$\overline{|Y|} = (2M/\pi)^{1/2}(1 + |E|^2_{\min}) \quad (26)$$

and

$$\sigma_Y = (1 + |E|^2_{\min})M^{1/2}(1 - 2/\pi)^{1/2}. \quad (27)$$

The departure of the mean value of the right-hand side of (24) from $|E(\mathbf{h})|$, if random values are used is,

Table 2. *A summary of results in applying SAYTAN to aPP data at different resolutions*

NREF is the number of reflections used, $|E|_{min}$ is the minimum $|E|$ used, NREL is the number of linking three-phase relationships and MPE is the mean phase error.

Resolution (Å)	NREF	$\|E\|_{min}$	NREL	Refinement process	Result [MPE (°)]	Minimum MPE (°)
1.0	800	1.7	9726	*SAYTAN*	11 sets [~40]	38
1.5	650	1.4	11620	*SAYTAN*	29 sets [~50]	48
1.77	556	1.3	14217	*SAYTAN*	16 sets [~55]	54
2.0	600	1.0	26323	*SAYTAN*	30 sets [~64]	62
2.25	350	1.2	6809	Parameter shift and *SAYTAN*	12 sets [~65]	63
2.5	300	1.14	5841	Parameter shift and *SAYTAN*	8 sets [~69]	68
3.0	315	0.9	9841	Parameter shift and *SAYTAN*	15 sets [~69]	69

Table 3. *Values of Q_Y for the SAYTAN trials described in Table 2*

	Resolution (Å)						
	1.0	1.5	1.77	2.0	2.25	2.5	3.0
Q_Y	0.90	0.73	0.91	0.85	0.44	0.34	0.34

as a number of standard deviations of the distribution,

$$Q_Y = 1.66[(M/N)^{1/2}(1 + |E|^2_{min})|E(\mathbf{h})| - (2/\pi)^{1/2}]. \quad (28)$$

When Table 1 was discussed previously it was noted that values of Q were quite small for low resolution and would probably mean that Sayre's equation could not be effectively used for finding phases. We shall now try to assess the significance of these values by calculating values of Q_Y for the results found by Mukherjee & Woolfson (1992).

The significance of Q values

It is possible from the information given in Table 2 to calculate values of Q_Y for any value of $|E(\mathbf{h})|$. The value of N is taken as 1200, which is approximate and does not include water, while M may be found from $6 \times$ NREL/NREF, remembering that each relationship contributes to three reflections and that a single contributor $E(\mathbf{k})E(\mathbf{h}-\mathbf{k})$ is partnered by another, which is $E(\mathbf{h}-\mathbf{k})E(\mathbf{k})$. The values of $|E|_{min}$ are also given at each resolution.

The values of Q_Y are given for the various resolutions in Table 3 for $|E(\mathbf{h})| = 1.4$, which we are estimating as an average for the larger E's in the system. It will be seen that these range from 0.91 to 0.34 and, coincidentally, the three lowest resolutions, with values of Q_Y from 0.34 to 0.44 did not give a solution with *SAYTAN* alone but had to be front-ended with a few cycles of a parameter-shift process.

On the basis of these results it is suggested that the value of Q_Y is a useful measure of the likelihood that an application of a Sayre-equation-based method would be successful and a lower value of Q_Y of about 0.35 is indicated for success. A comparison of Tables 2 and 3 also suggests that at any given resolution as large a system as possible should be used, although a lower limit of 3.5 Å resolution seems necessary for any method based on the Sayre equation even if all data is used.

Concluding remarks

Using the Sayre equation in a quantitative way offers advantages over the use of the more conventional tangent-formula method. However, there is a cost in computer time for the extra effectiveness although, for large structures, there is no alternative to paying this cost.

We have seen that there is an increase in power if more data are used and, ideally, it would be best to use all available reflections and relationships. This is not practicable for large structures for which the order of 10^4 reflections would require some 10^8 relationships to be considered. In this case there are advantages in changing from a reciprocal-space method to a real-space method and Zhang & Main (1990*a*,*b*) have successfully incorporated the Sayre equation into a real-space process of phase extension and refinement.

The analysis leading to the calculation of Q or Q_Y is only approximate and no allowance has been made for that aspect of using the Sayre equation in which a whole pattern of magnitudes is matched approximately rather than individual magnitudes precisely. Nevertheless, it is believed that the value of Q or Q_Y will indicate the possibility of success or otherwise in the application of the Sayre equation, say through *SAYTAN*, and finding a subset of E's and relationships which optimizes Q_Y could be a useful preliminary to actually running the program.

The question arises of whether the results given in this paper are typical of what might be expected with

proteins. The first point to note is that aPP contains a moderately heavy atom (zinc) and this certainly is helpful in solving the structure. If the zinc contribution is artificially subtracted from the observed structure factors, in a way which retains the errors of measurement, then no solution is found with the modified structure in 1000 trials with *SAYTAN*, even at 1.0 Å resolution. It is possible, even probable, that a larger number of trials would find a solution but, whatever the situation with respect to this particular structure, the principle of using a set of reflections and relationships giving as high as possible a value of Q_Y consistent with the need to minimize computing requirements to the capability of the available computer, should still be valid.

I am grateful to the Science and Engineering Research Council, the Wolfson Foundation and the Wellcome Trust for support of direct-methods work at York, of which this is a component.

References

HUGHES, E. W. (1957). *Acta Cryst.* **6**, 871.
KARLE, J. & HAUPTMAN, H. (1956). *Acta Cryst.* **9**, 635–651.
MUKHERJEE, M. & WOOLFSON, M. M. (1992). *Acta Cryst.* D**49**, 9–12.
SAYRE, D. (1952). *Acta Cryst.* **5**, 60–65.
SHIONO, M. & WOOLFSON, M. M. (1991). *Acta Cryst.* A**47**, 526–533.
WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
WOOLFSON, M. M. & YAO JIA-XING (1990). *Acta Cryst.* A**46**, 409–413.
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.