References

DEBAERDEMAEKER, T., TATE, C. & WOOLFSON, M. M. (1985). *Acta Cryst.* A**41**, 286-290.

DEBAERDEMAEKER, T., TATE, C. & WOOLFSON, M. M. (1988). *Acta Cryst.* A**44**, 353-357.

GLOVER, I., HANEEF, I., PITTS, J., WOOD, S., MOSS, T., TICKLE, I. & BLUNDELL, T. (1983). *Biopolymers*, **22**, 293-304.

SAYRE, D. (1952). *Acta Cryst.* **5**, 60-65.

WOOLFSON, M. M. (1958). *Acta Cryst.* **11**, 277-283, 393-397.

WOOLFSON, M. M. & YAO, J. (1990). *Acta Cryst.* A**46**, 409-413.

Acta Cryst. (1991). A**47**, 533-537

On the Application of One-Wavelength Anomalous Scattering. III. The Wilson-Distribution and MPS Methods

BY A. C. RALPH AND M. M. WOOLFSON

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

(*Received* 25 *June* 1990; *accepted* 25 *March* 1991)

Abstract

This paper describes two methods which break the ambiguity associated with phase determination from one-wavelength anomalous-scattering data when the positions of the anomalous scatterers are known. In the Wilson-distribution method the magnitudes of the contributions of the light atoms are found for each of the alternative phases and the phases are then given weights according to the usual Wilson probability distribution for the magnitudes. In the MPS method the two possible magnitudes of the contributions of the light atoms to the scattering are compared with a theoretical value based on the observed structure magnitudes and the Fourier coefficient of the $|P_s|$ function [Hao & Woolfson (1989). *Acta Cryst.* A**45**, 794-797]. Once again this leads to a weight for each alternative phase. A best-estimate phase based on the two weights is compared with true phases for two known proteins consisting of 36 and 96 amino acid residues respectively. It is concluded that the quality of the phase estimates is similar to that obtained by other previously published procedures and that the results are much more limited by the magnitude of the anomalous contribution and the data quality than by the actual method used. The methods were then applied to the smaller protein structure using calculated data both with and without added errors. It is concluded that this common procedure for the testing of methods must be done with great care, otherwise unduly optimistic conclusions may be drawn.

Introduction

We consider a structure in which there are m anomalous scatterers, whose positions are known, and n non-anomalous scatterers in the unit cell. From one-wavelength anomalous-scattering (OAS) data, with known positions for the anomalous scatterers, there will be two possibilities for the phase, $\varphi' \pm \Delta\varphi$, as shown in Fig. 1. The various quantities shown in Fig. 1 are related by

$$|F|^2 = \tfrac{1}{2}(|F^+|^2 + |F^-|^2) - |F''|^2 \qquad (1)$$

and

$$\cos(\Delta\varphi) = (|F^+|^2 - |F^-|^2)/2|F||F''|. \qquad (2)$$

There are various ways in which the ambiguity may be resolved or circumvented (see, for example, Okaya, Saito & Pepinsky, 1955; Kartha, 1961; Blow & Rossmann, 1961; Fan, Han, Qian & Yao, 1984). Other work, which is related to our approach but different in substance, has been done by Wang (1985), who has not so much resolved the ambiguity as solved structures *despite* the ambiguity by his solvent-flattening technique, and by Karle (1985) who has taken

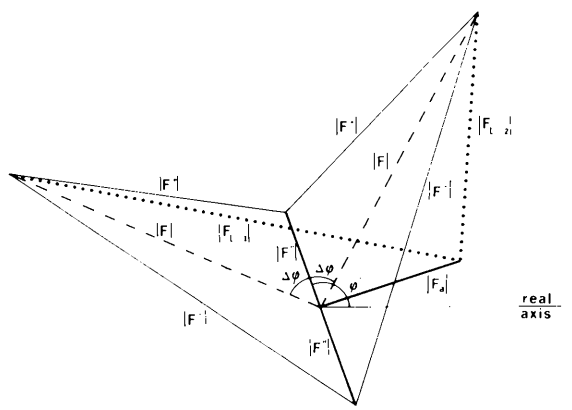

Fig. 1. The following contributions to the scattering are shown: $|F^+|$, $|F^-|$ the observed structure amplitudes of a Friedel pair; $|F''|$ the imaginary part of the contribution of the anomalous scatterers; $|F|$ the real part of the scattering from all scatterers; $|F_a|$ the total real part of scattering from the anomalous scatterers; $|F_{L1}|$, $|F_{L2}|$ the possible contributions of the light atoms.

0108-7673/91/050533-05$03.00

the positions of anomalous scatterers as unknown and has found phases relative to the phases of the contributions of the anomalous scatterers. The two new methods described here are related to the idea introduced by Kartha (1961) in which he suggested calculating a Fourier summation in which both possibilities for the phase were included. The density map which results is

$$\rho_b(\mathbf{r}) = V^{-1}\sum_{\mathbf{h}}\{|F(\mathbf{h})|[\cos\{2\pi\mathbf{h}\,.\,\mathbf{r}-(\varphi'+\Delta\varphi)\} + \cos\{2\pi\mathbf{h}\,.\,\mathbf{r}-(\varphi'-\Delta\varphi)\}]\}. \quad (3)$$

While both terms have equal weight, the correct set of terms should show the structure while the incorrect terms just contribute to a background noise. We have taken this idea one step forward by estimating weights to be associated with the alternative phases. Inclusion of these in the summation gives

$$\rho_w(\mathbf{r}) = V^{-1}\sum_{\mathbf{h}}\{|F(\mathbf{h})|[W^+\cos\{2\pi\mathbf{h}\,.\,\mathbf{r}-(\varphi'+\Delta\varphi)\} + W^-\cos\{2\pi\mathbf{h}\,.\,\mathbf{r}-(\varphi'-\Delta\varphi)\}]\} \quad (4)$$

where

$$W^+ + W^- = 1. \quad (5)$$

From (4) and (5) we find

$$\rho_w(\mathbf{r}) = V^{-1}\sum_{\mathbf{h}}|F(\mathbf{h})|[(1-2W^-)^2\sin^2(\Delta\varphi) + \cos^2(\Delta\varphi)]^{1/2}\cos\{2\pi\mathbf{h}\,.\,\mathbf{r}-(\varphi'+\delta)\} \quad (6)$$

where

$$\tan(\delta) = (1-2W^-)\sin(\Delta\varphi)/\cos(\Delta\varphi). \quad (7)$$

A sensible weighting scheme is one in which the signal-to-noise ratio is increased; the two methods we now describe differ only in the way the weights are calculated.

Theory

1. *The Wilson-distribution method*

It is evident from Fig. 1 that, after subtracting the scattering from all anomalous scatterers from the total scattering, there are two possible values for the magnitude of the structure factor, $|F_{L1}|$ and $|F_{L2}|$, for the remaining (and unknown) part of the structure. If the composition of the structure is known then Wilson statistics can be applied to the unknown part of the structure to give the relative probabilities of $|F_{L1}|$ and $|F_{L2}|$. The required Wilson distributions are

$$P(|F|) = (1/2\pi\Sigma)^{1/2}\exp(-|F|^2/2\Sigma) \quad (8)$$

for a centric structure and

$$P(|F|) = (2/\Sigma)|F|\exp(-|F|^2/\Sigma) \quad (9)$$

for an acentric structure, where

$$\Sigma = \sum_{j=1}^{N} f_j^2. \quad (10)$$

Therefore we can assign the following weights:

$$P(+\Delta\varphi) = W^+ = P(|F_{L1}|)/P(|F_{L1}|) + P(|F_{L2}|) \quad (11)$$

and

$$P(-\Delta\varphi) = W^- = P(|F_{L2}|)/P(|F_{L1}|) + P(|F_{L2}|). \quad (12)$$

2. *The MPS method*

This method estimates the magnitudes of the structure factors for the unknown part of the structure and compares the estimate with the values of $|F_{L1}|$ and $|F_{L2}|$. For a structure with only one type of anomalous scatterer, it is then possible to calculate an antisymmetric map, the P_s map (Okaya, Saito & Pepinsky, 1955; Hao & Woolfson, 1989), which shows positive peaks for vectors from anomalous scatterers to non-anomalous scatterers and negative peaks in the reverse direction. If we now consider the $|P_s|$ map, then this will give positive peaks for vectors in both directions and by Fourier transformation we can find the Fourier coefficients of this map, $\chi(\mathbf{h})$.

A Patterson map with Fourier coefficients $|F|^2$ (see Fig. 1) would show interatomic vectors between all atoms, while the Patterson map calculated with coefficients $|F_a|^2$ (see Fig. 1) would give vectors between anomalous scatterers only. It is clear that

vectors between non-anomalous scatterers only
= vectors between all atoms
− vectors between anomalous scatterers only
− vectors between anomalous and non-anomalous scatterers

which leads to

$$|F_L(\mathbf{h})|^2 = |F(\mathbf{h})|^2 - |F_a|^2 - k\chi(\mathbf{h}). \quad (13)$$

The constant k is a scale factor which will be discussed later. The only failure in this reasoning is that the P_s map, and therefore also the $|P_s|$ map, is affected by density cancellation where positive and negative peaks overlap so that (13) is not precisely true. However, disregarding this source of error, $|F_L|$ can be compared to $|F_{L1}|$ and $|F_{L2}|$ and weights deduced as follows, where it is assumed that $|F_{L1}| > |F_{L2}|$.

$$W^+ = (|F_L| - |F_{L2}|)/(2|F_L| - |F_{L1}| - |F_{L2}|) \quad \text{for } |F_{L2}| > |F_L| \text{ or } |F_L| > |F_{L1}|$$

$$W^+ = (|F_L| - |F_{L2}|)/(|F_{L1}| - |F_{L2}|) \quad \text{for } |F_{L1}| > |F_L| > |F_{L2}| \quad (14)$$

with $W^- = 1 - W^+$.

To obtain a valid estimate for $|F_L|$ the Fourier coefficients $\chi(\mathbf{h})$ need to be on the same scale as the observed data. Since the $|P_s|$ map is centrosymmetric

Table 1. *Details of test structures*

	APP	RNA
Space group	$C2$	$P2_12_12_1$
Number of molecules/ asymmetric unit	1	2
Number of amino acid residues/asymmetric unit	36	96
Type of anomalous scatterer	Hg	Pt
Effective number of asymmetric units/unit cell	2	6
Resolution of data (Å)	2·0	2·5
Number of independent reflections	2109	7009
Cell dimensions (Å, °) a	34·18	64·90
b	32·92	78·32
c	28·44	38·79
β	105·3	

then its Fourier coefficients are real and the scale factor may be calculated from

$$k^2 = \overline{\sum_{i=1}^{m}\sum_{j=1}^{n} f_j^2(f_i+f_i')^2}^{\mathbf{h}} \Big/ \overline{\chi(\mathbf{h})^2}^{\mathbf{h}} \qquad (15)$$

where f_i' is the real part of the anomalous scattering. This scale factor gives $k^2\overline{\chi^2}$ the correct theoretical value.

Results

Both methods were tested on two proteins, APP (Glover *et al.*, 1985) and RNA (Dodson, Sevcik, Dodson & Zelinka, 1987); details of these structures are given in Table 1. APP is a small protein with a strong anomalous signal from mercury. RNA is a larger protein with 20 anomalous-scatterer sites but with only partial occupancy so that there is roughly the equivalent of six Pt atoms per unit cell. The data for both structures are of good quality as judged by the normal standards of protein crystallography.

The results for both methods and for both structures with observed data are given in Table 2. The reflections are put in order of $|F|$ with the largest magnitude at the top. Not all the reflections are tabulated since, due to observational errors, some of them will have indicated negative $|F|^2$ [see (1)]. The cumulative mean phase error and the cumulative weighted mean phase error are given where the weights, R, associated with the overall phase estimates, $\varphi' + \delta$, are calculated from

$$R^2 = (1-2W^-)^2 \sin^2(\Delta\varphi) + \cos^2(\Delta\varphi). \qquad (16)$$

This weighting scheme has the reasonable property that if $\Delta\varphi = 0$ then any value of W^- will give $R = 1$ since there is no uncertainty in the phase. On the other hand, if $|\Delta\varphi|$ is large then the weight is a maximum for W^- equal to 0 or 1 which means that a definite choice of alternatives has been made. Finally it should be noted that $0 \le R \le 1$ for all values of $\Delta\varphi$ and W^-.

Both methods give similar results for APP and also for RNA. However, while the phases for APP could be used to produce an interpretable map (similar in quality to Figs. 1 and 2 of Hao & Woolfson, 1989), the same is not true for RNA. For the latter structure, with a weaker signal, the weights W^+ and W^- tend towards 0·5 - which is the Kartha approach. For a weak anomalous signal these methods add very little new information; this is also apparent in Table 2 where it can be seen that the weighted mean errors are very little different from the mean phase error for RNA.

Table 2. *Wilson-distribution method (left) and the MPS method (right) tested on observed APP and RNA data*

REFL: number of reflections in group; $|\bar{F}|$: mean $|F|$ [see (1)] in the range; ME: mean phase error (°); WME: weighted mean phase error (°).

REFL	$\|\bar{F}\|$	ME	WME	ME	WME
APP					
100	450	48·64	47·58	41·24	39·22
200	323	43·50	42·45	35·84	34·40
300	275	41·44	40·53	35·42	32·54
400	237	41·90	40·59	36·67	34·33
500	213	43·34	42·04	37·20	33·99
600	194	42·65	41·40	37·46	34·03
700	178	44·04	42·61	38·80	35·60
800	165	44·06	42·46	39·49	35·92
900	154	44·00	42·35	39·83	36·34
1000	141	43·88	42·30	40·68	36·65
1100	129	44·44	42·76	41·21	37·33
1200	120	44·18	42·51	41·30	37·16
1300	111	43·48	41·77	41·24	37·11
1400	103	43·64	41·67	42·32	38·30
1500	92	43·46	41·48	42·22	38·44
1600	83	43·77	41·50	42·19	38·56
1700	73	43·33	41·03	42·14	38·65
1800	62	43·51	41·05	42·43	39·32
1900	50	43·73	40·97	42·94	40·21
2000	35	43·79	40·81	43·20	40·21
2058	16	43·89	40·88	43·63	41·54
RNA					
500	750	56·55	54·76	63·72	61·25
1000	545	59·46	57·96	66·22	63·74
1500	460	60·37	59·07	66·00	63·16
2000	400	61·63	60·39	66·16	63·58
2500	354	63·25	62·06	66·97	64·79
3000	315	64·65	63·51	68·30	66·50
3500	279	65·52	64·48	69·34	67·60
4000	246	66·76	65·73	70·24	68·55
4500	216	67·87	66·80	70·94	69·75
5000	188	69·32	68·22	72·41	71·65
5500	160	70·31	69·15	73·82	73·14
6000	133	71·57	70·37	75·00	74·62
6500	102	72·63	71·35	76·11	76·06
7000	57	73·94	72·55	77·55	77·87
7002	6	73·94	72·56	77·55	77·87

Tests with calculated data

Very often a published account of a new method will illustrate its application using calculated data, sometimes with random errors superimposed to simulate real data. It occurred to us to test the validity of this approach and we decided to use the Wilson-distribution method and the structure APP for this purpose.

For the first trials we took (i) error-free calculated data and (ii) calculated data with a correct value for

Table 3. *Wilson-distribution method tested with calculated APP data (left) and calculated APP data with 10% error on* $\{|F^+|+|F^-|\}/2{\cdot}0$ *(right)*

REFL: number of reflections in group; $|\bar{F}|$: mean $|F|$ [see (1)] in the range; ME: mean phase error (°); WME: weighted mean phase error (°).

REFL	$\|\bar{F}\|$	ME	WME	$\|F\|$	ME	WME
100	554	13·87	12·36	562	23·28	22·30
200	360	13·23	11·23	365	23·40	22·18
300	304	12·07	10·27	306	22·81	21·62
400	270	10·45	8·83	268	22·46	21·27
500	245	12·18	10·16	241	22·26	20·97
600	222	11·05	9·22	219	22·89	21·45
700	203	11·44	9·39	205	23·33	21·70
800	188	11·72	9·57	187	24·33	22·49
900	173	11·83	9·64	174	24·09	22·27
1000	160	12·41	9·94	160	24·21	22·33
1100	146	12·57	10·06	149	24·19	22·29
1200	135	13·02	10·38	138	24·82	22·66
1300	125	13·66	10·91	127	25·59	23·37
1400	116	14·56	11·59	116	25·89	23·63
1500	106	15·02	11·93	107	26·72	24·34
1600	96	16·26	12·89	96	27·12	24·61
1700	84	17·56	13·82	85	27·74	25·10
1800	71	18·58	14·45	74	28·39	25·47
1900	59	19·41	14·97	61	29·09	25·99
2000	44	20·53	15·76	48	29·89	26·60
2100	24	22·68	17·18	27	31·65	27·78
2106	3	22·79	17·26	4	31·76	27·86

Table 4. *Wilson-distribution method tested on calculated APP data with separate errors applied to each intensity and background (see text)*

REFL: number of reflections in group; $|\bar{F}|$: mean $|F|$ [see (1)] in the range; ME: mean phase error (°); WME: weighted mean phase error (°).

REFL	$\|\bar{F}\|$	ME	WME
100	556	49·03	47·51
200	363	45·85	43·75
300	303	44·19	42·11
400	269	45·30	43·07
500	242	43·82	41·87
600	220	43·19	41·05
700	202	42·28	40·06
800	187	41·66	39·39
900	172	41·87	39·20
1000	159	41·36	38·72
1100	147	41·56	38·64
1200	135	41·03	38·13
1300	125	40·92	37·93
1400	115	40·87	37·87
1500	105	41·15	37·99
1600	95	41·76	38·34
1700	83	42·01	38·42
1800	71	42·27	38·48
1900	58	42·48	38·57
2000	44	42·55	38·64
2100	23	43·74	39·39
2103	4	43·74	39·40

the anomalous difference $\Delta F = |F^+| - |F^-|$ but with a random error applied to the mean magnitude $(|F^+| + |F^-|)/2$. The error was applied in the form of a factor ε drawn from a normal distribution of unit mean and standard deviation, σ, equal to 0·1. The results are shown in Table 3 and it is clear that they are both much better than those obtained from real data.

For our next trial we used a more complicated form of error based on the realistic scenario that both a peak height and a background are measured. We chose the same average background intensity, B, for all reflections and produced theoretical peak heights

$$|F_p^+|^2 = |F^+|^2 + B$$
$$|F_p^-|^2 = |F^-|^2 + B. \qquad (17)$$

The randomizing procedure previously described was used with a σ of 0·2 for B, $|F_p^+|^2$ and $|F_p^-|^2$, except for the centric reflections which were changed together to maintain equal magnitude. Error-transformed values for the structure amplitudes and the anomalous differences can now be obtained from

$$|F^+|' = \{|F_p^+|'^2 - B'\}^{1/2}$$
$$|F^-|' = \{|F_p^-|'^2 - B''\}^{1/2} \qquad (18)$$
$$\Delta F' = |F^+|' - |F^-|'$$

where B' and B'' are the two randomized values of B. The results of this calculation are shown in Table 4 and it will be seen that they are similar in overall quality to those obtained with the observed data.

Discussion

Two new methods have been explored for breaking the phase ambiguity which arises from the use of OAS data. The results for APP give a clear indication that the methods would probably work for small proteins with strong anomalous scatterers like mercury or platinum. However, we have included the results from the RNA data to show that these methods have quite distinct limitations. We believe that the limitation is not so much in the size of the structure but in the relative contributions to the intensity of anomalous and non-anomalous scattering. This can be defined as

$$\mathcal{R} = \sum_i \{(f_i + f_i')^2 + (f_i'')^2 - f_i^2\} / \sum_j f_j^2 \qquad (19)$$

where f, f' and f'' are the normal scattering factor and the real and imaginary parts of the anomalous contribution to the scattering factor respectively; the summation over i is for anomalous scatterers only and that over j is for all atoms. The value of $\mathcal{R}$ is about 0·023 for APP and 0·0016 for RNA, the latter value obviously being too low for either of our methods to give worthwhile results. An intermediate situation has been investigated by Karle (1985), who obtained good results from his method with calculated data and also calculated data with imposed errors for cytochrome c 550 with one Pt atom in the asymmetric unit, which gave a value of $\mathcal{R}$ about 0·008. We would estimate that for observed data a value of $\mathcal{R}$ below about 0·004–0·005 would not be expected to yield useful results for either of the methods we

describe here and possibly with the majority of other methods to be found in the literature.

The results are very similar to those for direct methods (Fan *et al.*, 1984), the P_s-function method (Hao & Woolfson, 1989) and the analytical method (Fan, Hao & Woolfson, 1990). This really raises the question whether the errors from the various methods are highly correlated with respect to individual reflections. If the correlation was low and the weighting schemes were sensible then it should be possible to produce a better answer than that from any single method by combining the results from all the methods. Our experiments in this direction show that it is indeed possible to produce a better result from pairs of methods, or even from triplets, but combining results from any more than three methods produces no gain. We suspect that each of the five methods we have explored in our laboratories in Beijing and York are exploiting the information from OAS in approximately equivalent ways and so are giving similar results. A related conclusion is that it would probably not be productive to look for even more methods of similar type for exploiting OAS data, although we do not discount the possibility that more sophisticated approaches may be more successful.

Finally, we point out from the results in Table 3 that it may be rather misleading to assess the practical effectiveness of procedures by using only calculated data sets - even those with added errors if the error simulation is not done realistically. The virtue of using calculated data is that they reveal the intrinsic properties of the method without the complication of errors in the data which will vary from one experimenter to another and with the technique of data collection. However, error-free data do not exist and errors do not occur in convenient ways. Methods like the P_s-function method, which use values of $|F(\mathbf{h})| - |F(\bar{\mathbf{h}})|$, can comfortably tolerate random errors of factors of, say, zero to three in these quantities. However, for real observed data, the error factors can be much higher and even the sign will be wrong for some of the anomalous differences. Whether or not any particular method will work with real data is not easy to predict. However, since there are very few occasions when real observed data cannot be used, we would advocate their use whenever possible. If this is not possible then, as can be seen from Table 4, it is possible to simulate errors in a more complicated way that gives data with characteristics comparable to those of observed data - including giving anomalous differences with the wrong sign.

We thank Hao Quan for providing the $\chi(\mathbf{h})$ Fourier coefficients from the $|P_s|$ map and also Eleanor Dodson for the FFT routine. Finally we are grateful to the Science and Engineering Research Council for its generous support.

References

BLOW, D. M. & ROSSMANN, M. G. (1961). *Acta Cryst.* **14**, 1195-1202.

DODSON, G. G., SEVCIK, J., DODSON, E. & ZELINKA, J. (1987). *Metabolism of Nucleic Acids Including Gene Manipulation*, pp. 33-36. Prague: Slovak Academy of Sciences.

FAN, H., HAN, F., QIAN, J. & YAO, J. (1984). *Acta Cryst.* A**40**, 489-495.

FAN, H., HAO, Q. & WOOLFSON, M. M. (1990). *Acta Cryst.* A**46**, 656-659.

GLOVER, I. D., MOSS, D. S., TICKLE, I. J., PITTS, J. E., HANEEF, I., WOOD, S. P. & BLUNDELL, T. L. (1985). *Adv. Biophys.* **20**, 1-12.

HAO, Q. & WOOLFSON, M. M. (1989). *Acta Cryst.* A**45**, 794-797.

KARLE, J. (1985). *Acta Cryst.* A**41**, 387-394.

KARTHA, G. (1961). *Acta Cryst.* **14**, 680-686.

OKAYA, Y., SAITO, Y. & PEPINSKY, R. (1955). *Phys. Rev.* **98**, 1857-1858.

WANG, B. C. (1985). *Diffraction Methods in Biological Macromolecules*, edited by H. W. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF, pp. 90-112. New York: Academic Press.

Acta Cryst. (1991). A**47**, 537-543

Some Additional Features of One-Wavelength Anomalous Dispersion

BY JEROME KARLE

Laboratory for the Structure of Matter, Naval Research Laboratory, Washington, DC 20375-5000, *USA*

(*Received* 9 *October* 1990; *accepted* 2 *April* 1991)

Abstract

By use of appropriate algebraic formulas, illustrations are given of several characteristics of one-wavelength anomalous-dispersion data, for the case that one predominant type of anomalous-scattering atom is present. It is shown that, when the structure of the anomalous scatterer is known, some simple algebraic formulas may be used to generate initial values of many phases associated with a macromolecular structure. In some cases, there may be enough phases determined to permit further refinement and

0108-7673/91/050537-07$03.00