Acta Cryst. (1997). D**53**, 673–681

New Techniques for Applying Anomalous-Scattering and Isomorphous-Replacement Data Incorporated in *ANOMIR* – a General Application Package

M. M. Woolfson,[a]* Yao Jia-Xing[a] and Fan Hai-Fu[b]

[a]*Department of Physics, University of York, York YO1 5DD, England, and* [b]*Institute of Physics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China. E-mail: mmw1@york.ac.uk*

(Received 14 October 1996; accepted 6 May 1997)

Abstract

A computer package *ANOMIR* is described which can derive phases from anomalous scattering and/or isomorphous-replacement data in any combination. For anomalous scattering it incorporates five methods of applying one-wavelength data and three methods for multiple-wavelength data including *SPIN*, reported here for the first time. In addition there are three procedures for multiple-wavelength data – the first modifying data for different wavelengths to make them mutually consistent, the second estimating the contributions of the anomalous scatterers alone and the third which finds anomalous differences. For single isomorphous replacement or one-wavelength anomalous scattering the phase ambiguity can be resolved by the direct method [Fan, Han, Qian & Yao (1984). *Acta Cryst.* A**40**, 489–495] but for multiple isomorphous replacement the main method is an adaptation of the probability-curve method [Blow & Crick (1959). *Acta Cryst.* **12**, 794–802]. A new statistical method is described for estimating the standard error in measuring magnitudes which is independent of having subsets of centric reflections. A method is described whereby the weights associated with phase estimates are used to generate probability curves, through which it is possible to combine estimates from different methods and to produce a 'best phase' and figure-of-merit for every reflection. *ANOMIR* procedures are also available for handling combinations of one-wavelength anomalous scattering with single- or multiple-isomorphous replacement. A final process, which is always beneficial, is a single parallel application of the tangent formula. The *ANOMIR* package has been designed for easy use and is controlled throughout by *KEYWORDS*. Results for several structures are given and compared with those found from the *MLPHARE* program in the *CCP*4 package.

1. Finding the anomalous scatterers

The first phasing of a protein structure solution is often carried out by the use of anomalous-scattering and/or isomorphous-replacement data. The data acquired can be of many types; for anomalous scattering it can be one-wavelength (OAS) or multiple-wavelength (MAS) and it can also be single isomorphous replacement (SIR) or multiple isomorphous replacement (MIR). A common procedure is where, for each available isomorph, OAS data is collected; if a single isomorph is available this gives SIROAS data, for many isomorphs MIROAS data.

For obtaining phases from either OAS or MAS data the first requirement is to determine the positions of the anomalous scatterers. This is commonly carried out by using the anomalous differences,

$$\Delta F = \big| |F(\mathbf{h})| - |F(\bar{\mathbf{h}})| \big|. \qquad (1)$$

For perfect data the magnitude of the imaginary part of the anomalous contribution of the anomalous scatterers, $|F''|$, must be greater than $\frac{1}{2}\Delta F$ (Fig. 1) so that when ΔF is large then the contribution of the anomalous scatterer to the structure factor must also be large, although unquantifiably so. This enables a subset of reflections to be found for which the anomalous contribution is large: it is only a subset because for small ΔF the contribution of the anomalous scatterers can be either large or small. To find the anomalous scatterers either the values of $(\Delta F)^2$ can be used as coefficients of a Patterson function, which should show vectors between the anomalous scatterers, or ΔF can be inserted as a structure amplitude in a direct-method procedure (Mukherjee, Helliwell & Main, 1989).

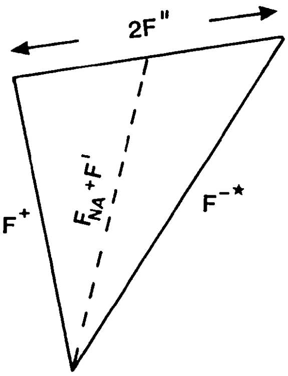


Fig. 1. The triangle formed by F^+, F^* and $2F''$.

ANOMIR contains the procedure *FDF* to find the anomalous differences and to put them in a form for subsequent use.

Where MAS data are available it is possible directly to find estimates of the magnitude of the contributions of both the anomalous scatterers, g, and of the total non-anomalous scattering, $|F_{\rm NA}|$ (Fig. 2), which is that of the protein plus the non-anomalous scattering of the anomalous scatterers. For each wavelength an equation can be derived linking the values of $|F_{\rm NA}|$ and g such that if $|F_{\rm NA}|$ is known then g can be found (Fan, Woolfson & Yao, 1993). This relationship is of the form,

$$Pg^4 + Qg^2 + R = 0, \qquad (2)$$

where P, Q and R involve known quantities, such as $|F(\mathbf{h})|$, $|F(\bar{\mathbf{h}})|$ and the real and imaginary parts of the anomalous component of the scattering factor for the anomalous scatterers f' and f''. They also involve the unknown quantity $|F_{\rm NA}|$ which is the contribution of the non-anomalous scattering to $|F(\mathbf{h})|$ and $|F(\bar{\mathbf{h}})|$. If data for several wavelengths are available then uniformly spaced values of $|F_{\rm NA}|$ are taken, spanning a range indicated by the values of $|F(\mathbf{h})|$ and $|F(\bar{\mathbf{h}})|$ and for each of them a value of g is calculated from (2). The quantity,

$$m = \frac{g}{\{(f')^2 + (f'')^2\}^{1/2}}, \qquad (3)$$

is the geometrical structure amplitude for the anomalous scatterers – that is the structure amplitude for a unit scattering factor, and should be independent of wavelength. The value of $|F_{\rm NA}|$ which gives the closest set of values of m for the different wavelengths is taken to indicate the best estimate of both $|F_{\rm NA}|$ and m, which is taken as the mean value of the closest set.

The procedure *FFM* in *ANOMIR* enables best estimates of values of m to be found. In general the use of values of m rather than ΔF leads to better results in determining the positions of the anomalous scatterers. As an example, in Fig. 3 we show Harker sections found both using anomalous differences and the values of m for the structure selenobiotinyl streptavidin (Hendrickson, Pähler, Smith, Satow, Merritt & Phizackerley, 1989). The space group is $I222$ with $a = 95.27$, $b = 105.40$, $c = 47.56$ Å, $Z = 8$. There are two independent Se atoms as anomalous scatterers and three sets of data were taken to 3.0 Å resolution. The anomalous-difference Harker section shown in Fig. 3(*a*) was produced with the best of the three data sets whereas values of m which gave the coefficients in Fig. 3(*b*) were derived by using all three sets of data. The much cleaner appearance of the m-derived map is evident and the superiority of using m values extends to situations where finding peaks is marginal and also to the accuracy of the peak coordinates.

When the positions of the anomalous scatterers have been found, either by the Patterson or direct methods, it

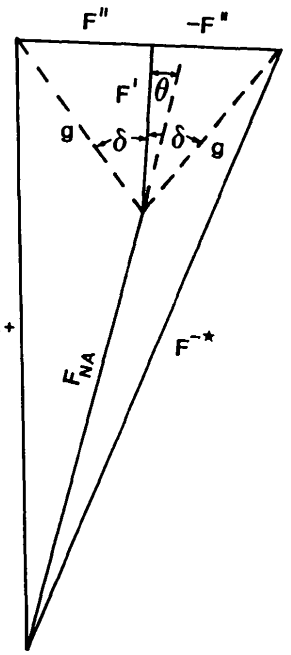
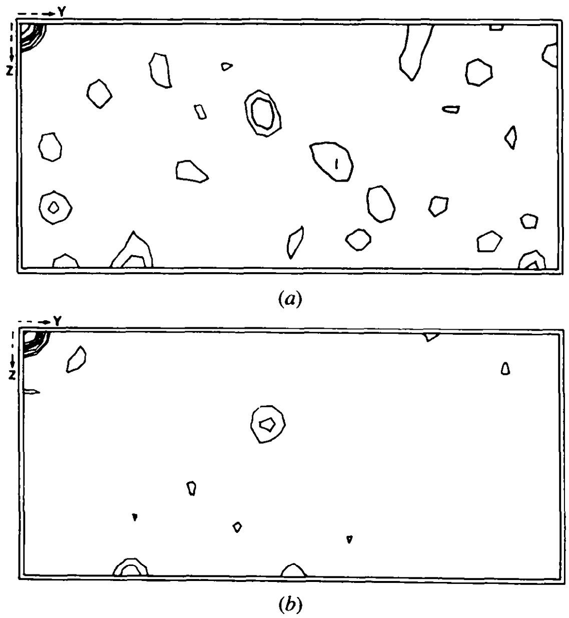


Fig. 2. Diagram illustrating the relationship between $F_{\rm NA}$, F' and F'' which contributes to the difference of the magnitudes $|F^+|$ and $|F^-|$. The quantity g is the magnitude of the anomalous contribution.



(*a*)

(*b*)

Fig. 3. Harker sections for selenobiotinyl steptavidin with (*a*) squares of anomalous differences as coefficients. The scale of the density is arbitrary but the contour levels have relative values 1, 2, 3, 4, 5 and 6 units. (*b*) Values of m^2 as coefficients. The scale of the density is arbitrary but the contour levels have relative values 1, 2, 3, 4, 5, 6 and 7.

is still necessary to determine the absolute configuration of the anomalous scatterers if the arrangement of anomalous scatterers is not centrosymmetric. The initially chosen configuration is automatically tested by means of a procedure described by Woolfson & Yao (1994) and, if necessary, it is changed to the other enantiomorph.

2. Finding phases with OAS data

Assuming that the positions of the anomalous scatterers can be determined with OAS data then inevitably there will be a phase ambiguity for each reflection (Fig. 4). Incorporated into *ANOMIR* there are five different techniques for handling OAS data which either resolve the ambiguity or circumvent it in some way by finding explicit phases. These are, the direct method (DM, Fan, Han, Qian & Yao, 1984); the P_S-function method (PS, Hao & Woolfson, 1989); the analytical method (AM; Fan, Hao & Woolfson, 1990); the Wilson-distribution method (WD; Ralph & Woolfson, 1991); and the modified P_s-function method (MPS; Ralph & Woolfson, 1991).

The P_S-function method was first described by Okaya, Saito & Pepinsky (1995). It leads to a final map which is, by the very nature of the method, a corrupted version of an electron-density map with uneven density at the sites of equal atoms and some atoms completely missing. The map is also particularly sensitive to data error since the original antisymmetric P_S-function map. $P_S(\mathbf{u})$, has coefficients which are the usually small differences of the anomalous intensities. However, the final map derived from the P_S function gives a specific phase estimate for each reflection, with a relative weight, $w_{\rm PS}$ which covers the range from 0 to 1 and depends on the product $|F_oF_c|$, where $|F_o|$ is the observed structure amplitude and F_c is the Fourier coefficient of the final map. This relative weight is found by dividing the reflections ranked by value of $|F_oF_c|$ into ten groups with equal numbers of reflections. For the top group the weights are between 1.0 and 0.9, for the next group between 0.9 and 0.8 and so on. Within each group the weight has linear dependence on the value of $|F_oF_c|$.

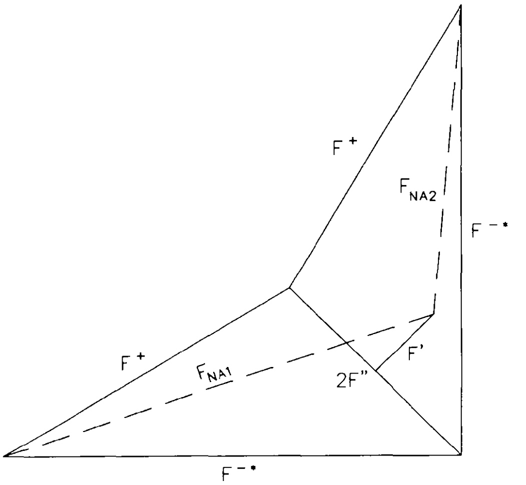


Fig. 4. The phase ambiguity with OAS data. Given $2F''$ in magnitude and phase there are two possible arrangements for F^+ and F^-. The respective contributions of the non-anomalous scattering are shown as $(F_{\rm NA})_1$ and $(F_{\rm NA})_2$.

The phase ambiguity for OAS appears in the form (Fig. 4),

$$\varphi(\mathbf{h}) = \varphi''_A(\mathbf{h}) \pm \Delta\varphi, \qquad (4)$$

where $\varphi''_A(\mathbf{h})$ is the phase of the imaginary contribution of the anomalous scattering and $\Delta\varphi$ comes from the known magnitudes of $F(\mathbf{h})$ and $F(\bar{\mathbf{h}})$ and the contributions of the anomalous scatterers. The direct method starts with the two possible phases from the ambiguity, giving each of them a weight equal to 0.5. In the original description of the direct method phase development was made through the use of the tangent formula which found a *best phase* with an associated weight. In *ANOMIR* the tangent-formula step is carried out by Fourier transformation of the squared density map, which is equivalent to a parallel application of the tangent formula in which new phases are found from previous phases simultaneously for all reflections.

The principle behind the P_S-function method shows that, with infinite resolution, if the positions of the anomalous scatterers are known then the phase ambiguity should not really be present. The analytical method uses some algebraic results plus the Fourier coefficients, $\chi(\mathbf{h})$, of the magnitude of the P_S function, $|P_S(\mathbf{u})|$, to derive phase estimates with associated weights.

The Wilson-distribution and MPS (modified P_S-function) methods both give weights, W^+ and W^- for the alternative phases given by (4), where $W^+. + W^- = 1$. The phase of the quantity,

$$R = W^- \exp\{i(\varphi''_A - \Delta\varphi)\} + (1 - W^-)\exp\{i(\varphi''_A + \Delta\varphi)\}, \qquad (5)$$

gives the estimated phase and the magnitude of R the associated weight. The alternative phases are associated with different contributions of the non-anomalous scatterers (Fig. 4); the Wilson distribution method uses theoretical intensity distributions (Wilson, 1949) to give relative probabilities to the two possible contributions and hence the associated phases. On the other hand the MPS method estimates the square of the magnitude of the contribution of the non-anomalous scatterers by considering Patterson coefficients of various vector sets, including $\chi(\mathbf{h})$, previously mentioned.

3. Finding phases with MAS data

The OAS methods can be applied to the individual data sets of MAS data but there are also three methods which use all the data sets together. However, where MAS data are available it is advantageous first to modify the data with a procedure called *REVISE* (Fan, Woolfson & Yao, 1993). It can be shown that the quantity,

$$C = \frac{|F(\bar{\mathbf{h}})|^2 - |F(\mathbf{h})|^2}{f''}, \tag{6}$$

should be independent of wavelength but, where several wavelengths are available, errors in that data show up in the non-equality of the values of C. *REVISE* is a procedure which obtains equality of the values of C by minimum modifications of the values of $|F(\mathbf{h})|$ and $|F(\bar{\mathbf{h}})|$ in terms of their standard deviations of measurement. It is found that better, sometimes much better, results are obtained from MAS methods with data modified by *REVISE*. The values of m used in Fig. 3(*b*) were obtained from data sets which had been through the *REVISE* process. This data, which came to us indirectly from Hendrickson's group, may have already been through a scaling procedure but nevertheless it still showed large inconsistencies in values of C and a better outcome was obtained after it was subjected to the *REVISE* procedure.

The method *AGREE*, which is specific to MAS, will be described in terms of the notation shown in Fig. 2. This differs from the previous description given by Fan, Woolfson & Yao (1993) although the general principle remains unchanged. For each wavelength an approximate value of $\sin\theta$ is found from,

$$\sin\theta = \frac{|F^-| - |F^+|}{2|F''|}. \tag{7}$$

If the positions of the anomalous scatterers have been found then the values of g can be calculated for all reflections. With the estimated value of $\sin\theta$ this can be used to find an estimate for $|F_{\rm NA}|$ from,

$$2|F_{\rm NA}|^2 + 4|F_{\rm NA}|g\cos\theta\cos\delta + 2g^2 - |F^-|^2 - |F^+|^2 = 0. \tag{8}$$

There is no ambiguity in the value of $|F_{\rm NA}|$ in general because of the quadratic nature of the equation since one of the solutions will be negative but there is an ambiguity in the value of $\cos\theta$ so two alternative estimates are found, $|F_{\rm NA}|_1$ and $|F_{\rm NA}|_2$. Each of these values is then used to get a revised estimate of $\sin\theta$ from,

$$\sin\theta = \frac{|F^-|^2 - |F^+|^2}{4|F_{\rm NA}|g\sin\delta}. \tag{9}$$

Iterative refinement for wavelength λ_i alternatively using (8) and (9) converges to the pairs of estimates $(|F_{\rm NA}|_{1,i}, \theta_{1,i})$ and $(|F_{\rm NA}|_{2i}, \theta_{2,i})$. Values found from the N different wavelengths are then tested for consistency for each alternative by,

$$T_j = \sum_{i=1}^{N} \sin\theta_{j,i} \quad B_j = \sum_{i=1}^{N} \cos\theta_j$$

$$\langle\theta_j\rangle = \arctan\left(\frac{T_j}{B_j}\right),$$

giving a figure-of-merit for the angle estimates for each alternative as,

$$\mathrm{FOM}(\theta_j) = 1.0 - \frac{T_j^2 + B_j^2}{N^2}. \tag{10a}$$

A figure of merit is also found for the values of $|F_{\rm NA}|$ from

$$\langle F_{{\rm NA},j}\rangle = \frac{1}{N}\sum_{i=1}^{N} |F_{\rm NA}|_{j,i} \quad \langle F^2_{{\rm NA},j}\rangle = \frac{1}{N}\sum_{i=1}^{N} |F_{\rm NA}|^2_{j,i}$$

and

$$\mathrm{FOM}(F_{{\rm NA},j}) = 1.0 - \frac{\langle F_{{\rm NA},j}\rangle^2}{\langle F^2_{{\rm NA},j}\rangle}. \tag{10b}$$

A combined figure of merit is found as,

$$(\mathrm{CFOM})_j = \mathrm{FOM}(\theta_j) + \mathrm{FOM}(F_{{\rm NA},j}) \tag{11}$$

and the smaller value of $(\mathrm{CFOM})_j$ is taken to indicate the true estimates of θ and $|F_{\rm NA}|$.

Another method specific to MAS is called *SPIN* which has replaced the method *ROTATE* described by Fan, Woolfson & Yao (1993). From Fig. 2 it can be seen that

$$|F^+|^2 = |F_{\rm NA}|^2 + g^2 + 2|F_{\rm NA}|g\cos(\theta + \delta) \tag{12}$$

and

$$|F^-|^2 = |F_{\rm NA}|^2 + g^2 + 2|F_{\rm NA}|g\cos(\theta - \delta). \tag{13}$$

For each wavelength, λ_i, measurements of $|F^+|$ and $|F^-|$ will be made with standard deviations σ_i^+ and σ_i^- and, since the two standard deviations are of similar magnitude, for a particular reflection we take the average standard deviation

$$\langle\sigma_i\rangle = \frac{1}{2}(\sigma_i^+ + \sigma_i^-). \tag{14}$$

If a case arises where the standard deviations σ_i^+ and σ_i^- are very different then it is better, from a theoretical point of view, to take the root-mean variance for $\langle\sigma_i\rangle$.

Since the positions of the anomalous scatterers are assumed to be known then the phase of g, φ_g, is known but the both the magnitude and phase of $F_{\rm NA}$, $|F_{\rm NA}|$ and $\theta_{\rm NA}$, are not known. The angle $\varphi_{\rm NA}$ is assigned values from 0 to 350° in steps of 10° and corresponding values of θ are found from

$$\theta = \varphi_g - \delta - \varphi_{NA}. \tag{15}$$

For each value of θ (12) and (13) are solved as quadratic equations in $|F_{NA}|$ to give $F_{1,i}$ and $F_{2,i}$ for wavelength λ_i. Since the equations are quadratic there are two pairs of values of $F_{1,i}$ and $F_{2,i}$ but the negative pair of values can be excluded. If the average of the $2k$ estimates of $|F_{NA}|$ is $\langle F\rangle$ then we assess the consistency of the $2k$ values, which we interpret as a probability, as the quantity

$$P(\varphi_F) = \exp\left[-\sum_{i=1}^{k}\frac{(F_{1,i}-\langle F\rangle)^2+(F_{2,i}-\langle F\rangle)^2}{2\langle\sigma_i\rangle^2}\right]. \tag{16}$$

These values of $P(\varphi_F)$ can be normalized to give unit area under the probability curve and then used either on its own or combined with other probability curves to give a *best phase* and a weight.

4. Isomorphous-replacement methods

Just as in the case of OAS, SIR gives a phase ambiguity. When the positions of the isomorphously replaced (i-r) atoms are known then their contributions are known in both magnitude and phase. What are also known are the structure amplitudes of the native protein, $|F_P|$, and of the derivative $|F_{PH}|$; the way in which the ambiguity arises is shown in Fig 5. The ambiguity may be expressed in the form,

$$\varphi_P = \varphi_H \pm \Delta\varphi, \tag{17}$$

which is similar to (4) and the direct method, as provided in *ANOMIR* can also be applied to SIR.

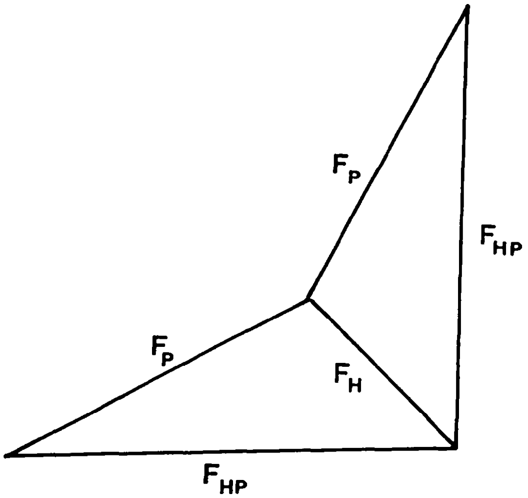


Fig. 5. The phase ambiguity for isomorphous replacement. If the contribution of the i-r atoms is known in magnitude and phase then there are two possible arrangements of the protein and derivative contributions, F_P and F_{PH}.

The basic process of MIR is that suggested by Blow & Crick (1959) which produces for each derivative a probability curve. Each curve is a sum of two Gaussian distributions, respectively centred on the alternative phases given by (17). To estimate the standard deviation of the distributions it is required to know the standard error in the measurement of the difference of the magnitudes $|F_P|$ and $|F_{PH}|$. Earlier ways of estimating this took advantage of centrosymmetric reflections where it is known that F_P and F_{PH} are likely to have the same sign. Estimation in this way depends on having a sufficient number of centric reflections to get a reasonable estimate in a statistical sense and, with a sufficient number of centrosymmetric reflections, the standard error can be determined in shells in reciprocal space. We now describe an alternative procedure for finding the standard deviation of the difference of $|F_P|$ and $|F_{PH}|$ which we denote by e.

5. Probability curves

The alternative procedure we have devised does not depend on the presence of centrosymmetric reflections and can also be applied to anomalous-scattering data. For simplicity we determined a single value of e for the whole of reciprocal space; numerical experiments in which e was calculated as a function of position in reciprocal space complicated the procedure and actually gave similar final outcomes. For example, in an early version of *ANOMIR* applied to the structure of RNASE (Sevcik, Dodson & Dodson, 1991) values of e dependent on $\sin\theta/\lambda$ gave a weighted mean phase error of 37.9° and a map correlation coefficient (MCC) of 0.516. With a uniform value of e over all reciprocal space the corresponding values were 38.5° with MCC = 0.525. We concluded that the differences are negligible and hence we have incorporated the determination of a single value of e in *ANOMIR*. We define

$$S = \left(\sum_{j=1}^{N} f^2_{\langle\theta\rangle}\right)^{1/2}, \tag{18}$$

where the scattering factors are taken at the mean scattering angle $\langle\theta\rangle$ for the data set and the summation is taken over all N atoms in the native protein. The quantity S^2 will be the mean intensity, without temperature effects, at the mean scattering angle and is a reasonable estimate of the mean intensity for the whole set. For the space group $P1$ a random normalized structure factor can be found by selecting a random number, r, with a uniform distribution between 0 and 1 and then taking

$$E = \left\{\ln\frac{1.0}{1.0-r}\right\}^{1/2}. \tag{19}$$

Values of E, so chosen, will have the distribution for $P1$ found by Wilson (1949). We now simulate a random protein structure factor by calculating

$$F_{\rm P} = SE\exp\left(-B\frac{\sin^2\langle\theta\rangle}{\lambda^2}\right). \qquad (20)$$

If the type and number of i-r atoms are known then an average value of the i-r atom contributions, $\langle F_{\rm H}\rangle$, can be found appropriate to the mean scattering angle $\langle\theta\rangle$.

We now take $F_{\rm P}$, $\langle F_{\rm H}\rangle$ and a random angle ψ which are combined to give a structure amplitude for the derivative from

$$|F_{\rm PH}|^2 = |F_{\rm P}|^2 + \langle F_{\rm H}\rangle^2 - 2|F_{\rm P}|\langle F_{\rm H}\rangle\cos\psi. \qquad (21)$$

In finding the standard error in the difference of the magnitudes the approach of Blow & Crick is followed by assuming that the errors of measurement are restricted to $F_{\rm PH}$. It is assumed that this error will have a Gaussian distribution with standard deviation $\sigma|F_{\rm PH}|$. A random error for $F_{\rm PH}$ is derived from s, a value selected from a Gaussian distribution with unit standard deviation, as

$$\Delta F = s\sigma|F_{\rm PH}|. \qquad (22)$$

A test is now made to see if it is possible to produce a triangle from the magnitudes of $|F_{\rm H}|$, $|F_{\rm P}|$ and $|F_{\rm PH}| + \Delta F$. If not, then a lack-of-closure is recorded. For a fixed value of σ the above procedure is carried out 10 000 times and the proportion of lack-of-closures, α, is recorded. This is then repeated for values of σ from 0.01 to 1.00 by steps of 0.01. For each trial the value of e^2 is taken as the mean of the values of $|\Delta F|^2$. While more elaborate procedures can be devised, for example using variable contributions for $F_{\rm H}$ or taking scattering-angle dependence into account, the simple procedure as described is found to give satisfactory results. In Blow and Crick's approach the error was assumed to reside only in $F_{\rm PH}$ but the variance was taken as $\sigma(|F_{\rm P}|)^2 + \sigma(|F_{\rm PH}|)^2$. Our approach has an equivalent effect in loading all the error on to $|F_{\rm PH}|$.

The same analysis can be carried out for anomalous scattering data where $2F''$ plays the role of $F_{\rm H}$. Table 1 shows partial results for the data at three separate wavelengths for selenobiotinyl streptavidin after the *REVISE* process had been applied.

Table 1. *Variation of lack of closure, α, and standard error in measuring $|F^+| - |F^-|$, e, with σ*

	$\lambda_1 = 0.9000$ Å		$\lambda_2 = 0.9795$ Å		$\lambda_3 = 0.9809$ Å	
σ	α	e	α	e	α	e
0.01	0.1764	7.0	0.2267	7.0	0.1677	7.0
0.02	0.2559	13.7	0.3427	13.7	0.2386	13.8
0.03	0.3341	20.7	0.4426	20.6	0.3126	20.7
0.04	0.3999	27.9	0.5204	27.8	0.3771	28.1
0.05	0.4451	35.2	0.5690	35.1	0.4230	35.3
0.06	0.4934	41.7	0.6168	41.6	0.4678	41.9
0.07	0.5397	48.4	0.6954	48.4	0.5133	48.7
0.08	0.5779	55.9	0.6899	55.8	0.5526	56.1
0.09	0.5987	62.0	0.7071	61.9	0.5725	62.3
0.10	0.6399	70.3	0.7395	70.2	0.6126	70.6
0.11	0.6554	76.1	0.7560	76.0	0.6302	76.5
0.12	0.6759	84.5	0.7767	84.4	0.6508	84.9
0.13	0.6926	90.2	0.7885	90.1	0.6692	90.7
0.14	0.7114	97.8	0.8016	97.6	0.6895	98.3

In fact after the *REVISE* process the columns of values of e should be very similar, as indeed they are. There are much greater differences if *REVISE* is not applied. The next step in the process is to find from the observed data the proportion of lack-of-closure situations and then by matching $\alpha_{\rm obs}$ to the table values an appropriate value of e can be chosen. Carrying out this process gives for the three wavelengths gives $\alpha_{\rm obs} = 0.3908$, 0.3866 and 0.3867 with corresponding values of $e = 27.0$, 17.1 and 29.7. These values of e are used in determining the standard deviations of the Gaussian functions used to produce the Blow & Crick probability curves. Here we have shown that probability curves can be generated both for anomalous scattering and isomorphous-replacement data. The probability curve method, *PROBABILITY*, in *ANOMIR* is the third of the methods applicable to MAS data and it is also applied to MIR data.

ANOMIR is not unique in using all reflections to determine the value of e or, more precisely, values of e which vary with $\sin\theta/\lambda$. In the *MLPHARE* procedure, which is part of the *CCP*4 package (Collaborative Computational Project, Number 4, 1994) only centric reflections are used initially, if available, in a cyclic procedure which determines not only the values of e but also includes the refinement of the occupancy and positions of the heavy atoms. In subsequent cycles all reflections, including the non-centric ones, are used - even if only the centric ones were used in the first cycle.

The use of probability curves, if available, provides an extremely effective way of combining results from different methods. Multiplying all curves together gives an overall probability curve $P(\Phi)$, which is first normalized and then from which the best phase is derived by

$$\tan\varphi_{\rm best} = \frac{\int P(\Phi)\sin\Phi{\rm d}\Phi}{\int P(\Phi)\cos\Phi{\rm d}\Phi} = \frac{T}{B}, \qquad (23a)$$

with a figure of merit, or weight, given by

$$w_{\rm best} = (T^2 + B^2)^{1/2}. \qquad (23b)$$

All the methods we have devised for OAS, the *AGREE* method for MAS and the direct method for SIR lead to either to a single phase estimate for each reflection accompanied by some weight or, where there remains a phase ambiguity, alternative phases with different weights. The weighting schemes tend to be relative within each method so that more confidence can be attached to a weight of 0.7 for the *AGREE* method than to the same weight for the P_S-function method. We have devised an empirical scheme based on our experiences with *ANOMIR* which enables probability curves to be derived from the weights for each particular method. It involves finding an equivalent standard deviations for the phase estimate, or phases estimates, for each reflection and then translating them into coefficients of a Cochran distribution (Cochran, 1955),

$$P(\varphi) = C\exp\{\kappa\cos(\varphi - \langle\varphi\rangle)\}, \qquad (24)$$

where C is a normalizing constant. We prefer the Cochran distribution to the Gaussian one as it has the proper periodicity for a function of angle. For a particular reflection and a particular method we first find the standard deviation, in degrees, as

$$\sigma = 104 - \sigma_m \times w, \qquad (25)$$

where w is the reflection weight and σ_m is 74 for the *AGREE* method and 34 for all other methods to which the process is applied. For the Cochran distribution there is a relationship between κ and variance or standard deviation (Karle & Karle, 1966) and we have devised a formula fit for the inverse transformation from σ to κ. Thus if, say, the Wilson distribution method gives the alternative phases as φ_1 with weight w_1 and φ_2 with weight w_2 then the corresponding σ_1 and σ_2 are found from (25), then transformed into corresponding values of κ_1 and κ_2 thus giving the probability curve as

$$P(\varphi) = C\{w_1\exp[\kappa_1\cos(\varphi - \varphi_i)] + w_2\exp[\kappa_2\cos(\varphi - \varphi_2)]\}, \qquad (26)$$

where C is an overall normalizing constant. If the method gives a single phase estimate and weight, as does the P_s-function method, then the weight is used to find the value of κ but does not appear as a multiplier of the Cochran distribution as in (26).

This way of producing equivalent probability curves is very effective for combining phase estimates. Poorly estimated phases give curves which are very flat and effect the phase estimate very little; well determined phases on the other hand give sharp peaks which dominate the phases estimate.

6. Some results

We now give results for a range of problems which illustrate the application of *ANOMIR* and we also offer comments on our experience in getting the best from the package. Run in its test mode *ANOMIR* gives not only unweighted mean phase errors (MPE) but also errors weighted with the value of $|F_{\rm obs}|$ (FMPE), errors weighted with the best weights (WMPE), and with the product of $|F_{\rm obs}|$ and the best weights (FWMPE). For FMPE and FWMPE the map correlation coefficients (MCC) are given for an unweighted map and a map calculated with Fourier coefficients weighted with the best weights. Another point to be noted is that *ANOMIR* detects reflections for which the data is very poor from the *REVISE* procedure and these are usually excluded from the phasing process.

6.1. *Selenobiotinyl streptavidin*

For this structure, containing 1984 atoms in the asymmetric unit, MAS data are available at three wavelengths. They were subjected to the *REVISE* procedure and the *SPIN* method was applied. It is our experience that with very good MAS data the application of *SPIN* alone often gives acceptable results. The quality of the data may be judged by the lack-of-closure errors for the data which have been given following Table 1. In the results that follow the mean phase errors in degrees are followed by the value of $\langle\cos\Delta\varphi\rangle$ or the MCC in parentheses. The results for 4435 reflections are, MPE 52.0°, FMPE 44.6° (0.554), FWMPE 38.1° (0.581), WMPE 41.7°. In the paper reporting the structure (Hendrickson, Pähler, Smith, Satow, Merritt & Phizackerley, 1989) in which the MAD technique was used the reported MPE was 56.9° but for 4598 reflections.

6.2. *RNASE (Sevcik, Dodson & Dodson, 1991)*

This structure has space group $P2_12_12_1$ with $a = 64.90$, $b = 78.32$, $c = 38.79$ Å and $Z = 4$. The asymmetric unit contains 1735 non-H atoms including water. There are three derivatives, containing Hg, Pt and I, and anomalous scattering data were taken for each derivative to 3.11, 2.50 and 2.52 Å resolution, respectively. The following methods were applied by *ANOMIR*: PS, AM, MPS, WD, DM and the *PROBABILITY* method which was applied to both MIR and OAS data. The outcome for 7054 reflections was, MPE 62.5°, FMPE 55.3° (0.469), FWMPE 50.6° (0.483), WMPE 54.3°.

We also applied the *MLPHARE* in the *CCP*4 package to this data with the following results for 7027 reflections, MPE 55.6°, FMPE 47.9° (0.525), FWMPE 39.6 (0.564), WMPE 46.4. The mean phase errors from *MLPHARE* are much less in all categories. Although the lack-of-closure errors for the anomalous scattering data were not too large, being 46.6, 19.7 and 24.4 for the three sets of data, it was decided to run *ANOMIR* with only the *PROBABILITY* procedure applied to the MIR and OAS data. The results for this were, MPE 55.8°, FMPE 47.2° (0.588), FWMPE 39.4° (0.641), WMPE 45.0°. This result is a distinct improvement on the *ANOMIR* result which used the additional procedures but similar to the *MLPHARE* results in MPE. However, it is much better than the *MLPHARE* result in terms of MCC for the weighted maps, which is the important characteristic for interpretation, presumably due to having a somewhat better weighting scheme. The OAS methods seem to have added noise rather than signal in conjunction with the probability-curve procedure.

6.3. *OPPAL (Glover, Denny, Nguti, McSweeney, Kinder, Thompson, Dodson, Wilkinson & Tame, 1995)*

This structure has space group $P2_12_12_1$ with $a = 110.50$, $b = 76.58$, $c = 70.67$ Å and $Z = 4$. The asymmetric unit contains 4662 non-H atoms including eight U atoms. Four-wavelength anomalous scattering data were taken to 2.2, 2.2, 2.1 and 2.3 Å resolution and the lack-of-closure errors, after the application of *REVISE* were 23.4, 47.0 42.9 and 28.8, respectively. These seemed low in relation to the expected magnitude of anomalous scattering from uranium so *ANOMIR* was run only with *SPIN*. The results for 30 527 reflections were, MPE 62.5°, FMPE 56.8° (0.462), FWMPE 51.9 (0.477), WMPE 54.7. The *MLPHARE* results, for 26 403 reflections, were better than this being, MPE 58.9°, FMPE 54.2° (0.494), FWMPE 46.0 (0.539), WMPE 50.1. We repeated the *ANOMIR* run with the addition of PS, AM, WD, DM and *AGREE* to *SPIN*. The results now were MPE 58.9°, FMPE 52.5° (0.533), FWMPE 47.5 (0.545), WMPE 51.1. These results are better than for *ANOMIR* with *SPIN* only and very marginally better than *MLPHARE* in terms of MCC but, taken in conjunction with the RNASE result, it raises the question of how to predict in advance what is the best approach. So far an infallible recipe for making such a decision has eluded us but *ANOMIR* is a time-efficient procedure so running it twice or even more times with different sets of methods to see which gives the most promising initial map is quite feasible.

This structure gave a convincing illustration of the effectiveness of the *REVISE* procedure in the application of FFM to the determination of the positions of the eight U atoms. When direct methods were used with the values of m found by the procedure FFM without *REVISE* only six peaks were found within 1 Å of the correct positions and these were in positions 1, 2, 3, 6, 7 and 8 in the list ordered by peak height. When *REVISE* was applied to the data the top eight peaks were all U atoms and the mean error in their positions was 0.42 Å. There was also a considerable fall, 616 to 471, between the heights of the eighth and ninth peaks.

6.4. *UTPASE (Cedergren-Zeppezauer, Larsson, Nyman, Dauter & Wilson, 1992)*

For this structure the space group is $R3$ with $a = 86.64$, $c = 62.23$ Å and $Z = 9$. The asymmetric unit contains 1028 non-H atoms plus 183 water molecules. Data is available from the native protein to 1.90 Å resolution and anomalous data from two derivatives, one containing Hg (to 1.99 Å resolution) and the other Pt (to 2.11 Å resolution). The lack-of-closure errors for the anomalous scattering data were extremely high, 296.6 and 306.2, so we ran *ANOMIR* with only the *PROBABILITY* procedure applied to the MIR and OAS data. This gave, for 11 707 reflections, MPE 52.8°, FMPE 43.7° (0.667), FWMPE 38.6 (0.688), WMPE 45.3.

The results from *MLPHARE*, for 11 687 reflections, were, MPE 59.0°, FMPE 50.7° (0.598), FWMPE 41.2 (0.641), WMPE 47.6. The results were appreciably better than those from *MLPHARE* in this case.

7. Conclusions

The *ANOMIR* package includes a number of features which are not available elsewhere and three of which – the *SPIN* method, the statistical approach to deriving lack-of-closure errors and the transforming of weights to Cochran-style probability curves – have been described here for the first time. The results it gives are generally similar to those from *MLPHARE*, tending to be somewhat better if the right combination of procedures can be found. However, the RNASE and UTPASE examples suggest that occasionally appreciably better results may be found with *ANOMIR*. The applications made with *ANOMIR* are still limited in number but it is hoped that as more experiments are performed, and as it is used to solve unknown structures, so a body of experience will build up leading to its more efficient use. Our advice is to run *ANOMIR* with two or three different combinations of methods and to accept that set of phases appearing to give a better map.

Comparisons with *MLPHARE* are really difficult to make since a final process in *ANOMIR* is to apply the tangent formula which may be considered as the first step of a refinement process.

In practice before resorting to maps, phase sets from any source should be subjected to some phase extension and refinement process. All the initial phase sets from the structures used in our trials have been subjected to the phase extension and refinement program *PERP* (Refaat, Tate & Woolfson, 1996) with the following eventual outcomes.

Selenobiotinyl streptavidin: this required only phase refinement. The final MPE was 50.9°, FWMPE 42.0° with MCC 0.710.

RNASE: phase information was extended to 1.8 Å resolution. The final MPE was 51.6°, FWMP 43.9° with MCC 0.693.

OPPAL: this required only phase refinement. The final MPE was 49.8°, FWMPE 41.5° with MCC 0.739.

UTPASE: phase information was extended to 1.89 Å resolution. The final MPE was 36.9°, FWMPE 28.9° with MCC 0.856.

It is clear that all these final phase sets, which have been obtained directly from the data without any structural knowledge being employed, should all be good starting points for elucidating the structures.

Experience in other areas of crystallography, for example in the application of direct methods, shows that it can be useful to have alternative procedures available. While all the procedures may be successful in the majority of cases there may be particular circumstances where one or other of the methods gives an advantage. What *ANOMIR* offers uniquely is an easy-to-use comprehensive set of techniques for handling OAS data which can be very useful in some circumstances. No test has been made of its limitations with respect to the size of structure being investigated but this would certainly depend on the values of the contributions of the anomalous scattering or i-r atoms relative to those of the non-anomalous scattering or native protein in conjunction with the quality of the data. Such considerations apply to all methods and not just those in the *ANOMIR* package.

The program systems *ANOMIR* and *PERP* are available by application to one of us (MMW, e-mail: mmwl@york.ac.uk).

We take this opportunity to thank the Leverhulme Trust for its generous support of this project.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.

Cedergren-Zeppezauer, E. S., Larsson, G., Nyman, P. O., Dauter, Z. & Wilson, K. S. (1992). *Nature (London)*, **335**, 740-743.

Cochran, W. (1955). *Acta Cryst.* **8**, 473-478.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760-763.

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**, 659-664.

Fan, H., Woolfson, M. M. & Yao, J. (1993). *Proc. R. Soc. London Ser. A*, **442**, 13-32.

Glover, I. D., Denny, R. C., Nguti, N. D., McSweeney, S. M., Kinder, S. H., Thompson, A. W., Dodson, E. J., Wilkinson, A. J. & Tame, J. R. H. (1995). *Acta Cryst.* D**51**, 39-47.

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

Hendrickson, W. A., Pähler, A., Smith, J. L., Satow, Y., Merrit, E. A. & Phizackerley, R. P. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 2190-2194.

Karle, J. & Karle, I. L. (1966). *Acta Cryst.* **21**, 849-859.

Mukherjee, A. K., Helliwell, J. R. & Main, P. (1989). *Acta Cryst.* A**45**, 715-718.

Okaya, Y., Saito, Y. & Pepinsky, R. (1955). *Phys. Rev.* **98**, 1857-1858.

Ralph, A. C. & Woolfson, M. M. (1991). *Acta Cryst.* A**47**, 533-537.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1996). *Acta Cryst.* D**52**, 1119-1124.

Sevcik, J., Dodson, E. J. & Dodson, G. G. (1991). *Acta Cryst.* B**47**, 240-253.

Wilson, A. J. C. (1949). *Acta Cryst.* **8**, 318-321.

Woolfson, M. M. & Yao, J. (1994). *Acta Cryst.* D**50**, 7-10.