Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

SAD or MAD phasing: location of the anomalous scatterers

Maria Cristina Burla,[a] Benedetta Carrozzini,[b] Giovanni Luca Cascarano,[b] Carmelo Giacovazzo[b,c]* and Giampiero Polidori[a]

[a]Dipartimento di Scienze della Terra, Piazza Università, 06100 Perugia, Italy, [b]Istituto di Cristallografia, CNR, c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and [c]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy

Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

Received 2 December 2002
Accepted 23 January 2003

The method of joint probability distribution functions is applied in order to estimate the structure-factor moduli of the anomalous scatterer substructure both in the SAD (single-wavelength anomalous dispersion) and in the MAD (multi-wavelength anomalous dispersion) cases. The experimental data $|F_1^+|$, $|F_1^-|$, ..., $|F_n^+|$, $|F_n^-|$ measured at n wavelengths are used simultaneously to estimate the value of $|F_{oa}|$ arising from the normal scattering of the anomalous scatterers. A practical procedure is described that, when applied to the experimental diffraction data of several proteins, shows robustness and efficiency.

1. Notation

N: number of atoms in the unit cell.
a: number of anomalous scatterers in the unit cell.
$na = N - a$: number of non-anomalous scatterers.
$f_j = f_j^0 + \Delta f_j + if_j'' = f_j' + if_j''$: scattering factor of the jth atom. f' is its real and f'' its imaginary part. The thermal factor is included.
$\Sigma_{Np} = \sum_{j=1}^{N}(f_j'^2 + f_j''^2)$. The summation is calculated at the pth wavelength and is extended to all atoms in the unit cell.
$\Sigma_o = \sum_{j=1}^{na}(f_j^o)^2$. The summation is extended to all non-anomalous scatterers in the unit cell.
$\Sigma_{oa} = \sum_{j=1}^{a}(f_j^o)^2$. The summation is extended to all the anomalous scatterers in the unit cell.
$F^+ = |F^+|\exp(i\varphi^+) = F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(2\pi i\mathbf{hr}_j)$.
$E^+ = F^+/(\varepsilon\Sigma_N)^{1/2} = R\exp(i\varphi^+) = A^+ + iB^+$.
$F^- = |F^-|\exp(i\varphi^-) = F_{-\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(-2\pi i\mathbf{hr}_j)$.
$E^- = F^-/(\varepsilon\Sigma_N)^{1/2} = G\exp(i\varphi^-) = A^- + iB^-$.
F_p^+, F_p^-, $E_p^+ = A_p^+ + iB_p^+$, $E_p^- = A_p^- + iB_p^-$ denote the values for the pth wavelength.
n: number of wavelengths.
$F_{oa} = |F_{oa}|\exp(i\varphi_{oa}) = \sum_{j=1}^{a} f_j^o \exp(2\pi i\mathbf{hr}_j)$.
$E_{oa} = F_{oa}/(\varepsilon\Sigma_{oa})^{1/2} = R_{oa}\exp(i\varphi_{oa}) = A_{oa} + iB_{oa}$.
$\Delta_{\rm ano} = |F^+| - |F^-|$.
The paper by Burla *et al.* (2002) will be referred to as paper I.

2. Introduction

The roles of the SAD (single-wavelength anomalous scattering) and MAD (multi-wavelength anomalous dispersion) techniques have notably increased in the last few years as a consequence of the tunability of synchrotron radiation. If the data are accurately measured, the crystal structure may be solved by single-wavelength anomalous diffraction (Hendrickson & Teeter, 1981; Wang, 1985; Dauter *et al.*, 2002). The MAD technique requires more experimental engagement but may provide more accurate phase estimates.

Table 1
Set of test structures.

n_{wl} is the number of wavelengths used in the experiment, An. scat. gives the atomic species of the anomalous scatterers, *na* is the number of anomalous scatterers and Resol. is the limiting resolution to which the data are measured.

Protein code	Space group	n_{wl}	An. scat.	*na*	Resol. (Å)	Reference
ApD	$C222_1$	4	Se	3	2.2	Walsh *et al.* (1999)
JIA	$C222_1$	4	Se	8	2.5	Li *et al.* (2000)
KPR	$P4_22_12$	3	Se	8	2.3	Matak-Vinkovic *et al.* (2001)
PSCP	$P6_2$	3	Br	13	1.8	Dauter *et al.* (2001)
Cyanase	P1	4	Se	40	2.4	Walsh *et al.* (2000)
Tm0665	$P2_1$	3	Se	45	2.0	Lesley *et al.* (2002)
TGEV	$P2_1$	4	Se	60	2.9	Anand *et al.* (2002)
AEP	$P2_1$	3	Se	66	2.55	Chen *et al.* (2000)
Glucose isomerase	$I222$	1	Mn	1	1.5	Dauter *et al.* (2002)
2Zn insulin	$R3$	1	Zn	2	1.0	Dauter *et al.* (2002)
Ca subtilisin	$P2_12_12_1$	1	Ca	3	1.75	Dauter *et al.* (2002)
CauFd	$P4_32_12$	1	Fe	8	0.94	Dauter *et al.* (2002)
DNA	$P2_12_12_1$	1	P	10	1.5	Dauter *et al.* (2002)
Lysozyme	$P4_32_12$	1	S, Cl	10 + 7	1.53	Dauter *et al.* (2002)
CUTA1	$P2_12_12_1$	1	Hg	18	2.5	Calderone *et al.* (2002)
APT	$P2_1$	1	Br	22	1.8	Dauter *et al.* (2002)

Two procedures can in principle be used for determining phases *via* SAD or MAD.

(i) Use of the triplet invariant estimates given six diffraction magnitudes (Hauptman, 1982; Giacovazzo, 1983). The protein phases are directly derived *via* a tangent formula without any prior knowledge of the anomalous scatterer positions. The experimental success of this approach has so far been modest and may only be used for the SAD case.

(ii) The two-step technique, in which the structural parameters of the anomalous scatterers are first determined and refined and then, in the second step, the protein phases are assigned. The technique has been described by several authors (Karle, 1980; Hendrickson, 1985; Pähler *et al.*, 1990; Terwilliger, 1994) and involves the interpretation of the Patterson function (Sheldrick *et al.*, 1993; Sheldrick, 1998; Terwilliger & Berendzen, 1999; Grosse-Kunstleve & Brunger, 1999). It constituted a standard in the field until a few years ago, when second-generation direct-methods programs (Miller *et al.*, 1994; Sheldrick, 1998; Burla *et al.*, 2001; Foadi *et al.*, 2000) erupted into the area of macromolecular crystallography. The use of such programs was encouraged by the practice of introducing Se atoms into a protein as selenomethionines: indeed, the number of anomalous scatterers may be quite large, sometimes greater than 50.

Shake-and-Bake (Smith *et al.*, 1998; Howell *et al.*, 2000), as well as *SHELXD* (Schneider & Sheldrick, 2002), derive the coordinates of the anomalous substructure from a single wavelength: the other wavelengths, when available, are used to identify and eventually confirm the correct solution. A new approach was suggested in paper I, in which the authors applied the rigorous method of joint probability distribution functions to estimate the amplitudes of the structure factors of the anomalously scattering substructure given the experimental diffraction moduli. The method was restricted to two wavelengths: its advantage is that the estimates can simultaneously exploit the anomalous and the dispersive differences. The first applications were very encouraging.

This paper is devoted to the following.

(*a*) Generalizing the method suggested in paper I to the *n*-wavelength case (including the $n = 1$ case).

(*b*) Defining a robust procedure for finding the anomalously scattering substructure. The approach does not use the dual-space recycling techniques used by *Shake-and-Bake* and *SHELXD*, but employs the tangent formula as the first step and real-space techniques in the next steps, as suggested in the paper by Burla *et al.* (2003). Furthermore, no use of the Patterson function is made, as is performed in *SHELXD*.

(*c*) Describing the applications to an extended set of experimental data (see Table 1, where the main crystallochemical properties of the test structures are given).

3. The joint probability distribution function $P(R_{oa} \mid R_1, \ldots, R_n, G_1, \ldots, G_n)$

As in paper I, the positions of all the atoms in the asymmetric unit will be the primitive random variables of our probabilistic approach. Furthermore,

$$F_j^+ = F_{aj}^+ + F_{naj}^+ + |\mu_j^+| \exp(i\theta_j^+)$$
$$F_j^- = F_{aj}^- + F_{naj}^- + |\mu_j^-| \exp(i\theta_j^-), \quad j = 1, ..., n,$$

where μ^+ and μ^- are the measurement errors relative to F^+ and F^-, respectively; they will be treated as additional primitive random variables.

In paper I, the $n = 2$ wavelength case was studied. The joint probability distribution

$$P_2 = P(E_{oa}, E_1^+, E_1^-, E_2^+, E_2^-)$$

was derived in the form of the ten-dimensional $(4n + 2)$ Gaussian distribution [see equation (3) in paper I],

$$P_2 = P(A_{oa}, A_1^+, A_2^+, A_1^-, A_2^-, B_{oa}, B_1^+, B_2^+, B_1^-, B_2^-)$$
$$= \pi^{-5}(\det \mathbf{K})^{1/2} \exp(-\tfrac{1}{2}\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T}), \quad (1)$$

where $\mathbf{K}$ is a symmetric variance–covariance square matrix, $\mathbf{K}^{-1} = \{\lambda_{ij}\}$ is its inverse and $\mathbf{T}$ is a suitable vector with components defined in terms of the variables $A_{oa}, A_1^+, A_2^+, A_1^-, A_2^-, B_{oa}, B_1^+, B_2^+, B_1^-, B_2^-$.

The same mathematical technique may be used for the study of the distribution

$$P_n = P(A_{oa}, A_1^+, A_2^+, ..., A_n^+, A_1^-, A_2^-, ..., A_n^-, B_{oa}, B_1^+, B_2^+, ..., B_n^+, B_1^-, B_2^-, ..., B_n^-).$$

We obtained the joint probability density (details are not given for brevity)

$$P_n = \pi^{-(2n+1)}(\det \mathbf{K})^{1/2} \exp(-\tfrac{1}{2}\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T}), \quad (2)$$

where $\mathbf{K}$ is a symmetric square matrix of order $(4n + 2)$, $\mathbf{K}^{-1} = \{\lambda_{ij}\}$ is its inverse and $\mathbf{T}$ is a suitable vector with components defined in terms of the variables $A_{oa}, A_1^+, A_2^+, \ldots, B_n^-$.

In paper I we explicitly defined, for $n = 2$, the algebraic expression of all the elements of the matrix $\mathbf{K}$: since $\mathbf{K}$ is of

order 10, we derived the expressions of $n_e = (9 \times 10)/2 + 10 = 55$ entries. For any value of n, we should now render explicit the values of

$$n_e = [(4n+1)(4n+2)/2 + (4n+2)] = (2n+1)(4n+3)$$

entries. We see that $n_e = 105$ for $n = 3$, $n_e = 171$ for $n = 4$, $n_e = 253$ for $n = 5$ and so on. The most useful way of dealing with this problem is to define an algorithm for deriving the expressions of the entries of **K** for any value of n rather than giving them explicitly. Such an algorithm is described in Appendix *A*. It is also shown in Appendix *A* that the association of the various components of **T** with the various λ_{ij} is trivial. Accordingly, the full distribution (2) has been algorithmically constructed in our computer program; it is possible to use up to five wavelengths (the $n = 1$ case included).

The next steps for obtaining the estimate of R_{oa} are as follows.

(i) Change (2) into

$$P(R_{oa}, R_1, ..., R_n, G_1, ..., G_n, \varphi_{oa}, \varphi_1^+, ..., \varphi_n^+, \varphi_1^+, \varphi_1^-, ..., \varphi_n^-) \quad (3)$$

via the change of variables

$$\begin{aligned} A_{oa} &= R_{oa}\cos\varphi_{oa}, \quad & B_{oa} &= R_{oa}\sin\varphi_{oa}, \\ A_j^+ &= R_j\cos\varphi_j^+, \quad & B_j^+ &= R_j\sin\varphi_j^+, \\ A_j^- &= G_j\cos\varphi_j^-, \quad & B_j^- &= G_j\sin\varphi_j^-. \end{aligned}$$

(ii) Integrate (3) over the phase variables and define the conditional distribution

$$P(R_{oa}|R_1, ..., R_n, G_1, ..., G_n).$$

(iii) Calculate the expected value

$$\langle R_{oa}|R_1, ..., R_n, G_1, ..., G_n\rangle.$$

At the end of this process, we obtain

$$\langle R_{oa}|R_1, ..., G_n\rangle = 2^{-1}(\pi/\lambda_{11})_1F_1(-1/2; 1; -X^2/\lambda_{11}), \quad (4)$$

where ${}_1F_1$ is the confluent hypergeometric function,

$$\begin{aligned} X^2 &= Q_1^2 + Q_2^2, \\ Q_1 &= \lambda_{12}R_1 + \lambda_{13}R_2 + ... + \lambda_{1,n+1}R_n + \lambda_{1,n+2}G_1 + ... \\ &\quad + \lambda_{1,2n+1}G_n, \\ Q_2 &= \lambda_{1,2n+3}R_1 + \lambda_{1,2n+4}R_2 + ... + \lambda_{1,3n+2}R_n + ... - \lambda_{1,3n+3}G_1 \\ &\quad - ... - \lambda_{1,4n+2}G_n. \end{aligned}$$

Since ${}_1F_1(-1/2; 1; -z^2)$ is well approximated (see paper I) by the hyperbole $y = (1 + 2z^2/\pi^{1/2})^{1/2}$ in the full range $(0, \infty)$, the expected value of R_{oa} may be calculated *via* the simpler expression

$$\langle R_{oa}|R_1, ..., G_n\rangle = \tfrac{1}{2}(\pi/\lambda_{11})^{1/2}[1 + 4X^2/(\pi\lambda_{11})]^{1/2}. \quad (5)$$

The standard deviation of the estimate is calculated as described in paper I,

$$\sigma_{R_{oa}} = (\langle R_{oa}^2|...\rangle - \langle R_{oa}|...\rangle^2)^{1/2} = \left[\left(1 - \frac{\pi}{4}\right)\lambda_{11}^{-1}\right]^{1/2},$$

from which

$$\frac{\langle R_{oa}|...\rangle}{\sigma_{R_{oa}}} = \left[\frac{(\pi/4) + (X^2)/\lambda_{11}}{1 - (\pi/4)}\right]^{1/2}. \quad (6)$$

4. The accuracy of the $\langle R_{oa}|\ R_1, \ldots, G_n\rangle$ values

Once the moduli R_{oa} have been estimated, a direct phasing procedure may be applied (see §5); its efficiency depends on some critical steps connected to the use of (5) and (6). We briefly discuss these below.

(i) *Is MAD always preferable to SAD?* In an ideal (non-realistic) situation the answer is trivial: MAD should be preferred because SAD estimates are intrinsically ambiguous and not available for symmetry-restricted reflections. In practice, however, experimental errors can question the primacy of the MAD techniques, particularly when SAD data present a high redundancy of the measurements. We will show some examples in §6.

(ii) *Are the MAD estimates for the n-wavelength case more accurate than for the (n − 1) or the (n − 2) wavelength cases?* To answer this question, let us examine the intuitive general conditions for obtaining good estimates in the two-wavelength case: (*a*) if f_1'' and f_2'' are both sufficiently large they intrinsically secure good SAD estimates and (*b*) if, in addition, $|\Delta f_1'|$ and $|\Delta f_2'|$ are also both large, they can efficiently correct the twofold ambiguity of the SAD estimates. Such conditions, however, do not take into full consideration the properties of the **K** matrix: in the case of small measurement errors, if $f_1'' \simeq f_2''$ and $\Delta f_1' \simeq \Delta f_2'$, two columns of **K** tend to be identical. det(**K**) then approaches zero, the problem becomes ill-conditioned and the resulting phase estimates tend to be unreliable. The same will occur in the *n*-wavelength case if one column is the linear combination of other two columns. Typical results are shown in Table 2 (columns 2 and 4), where we have applied (5) to the calculated data of two test structures. The accuracy of the estimates is measured by the parameter

$$R_A = \frac{\sum|(R_{oa})_t - S\langle R_{oa}|....\rangle|}{\sum(R_{oa})_t},$$

where $(R_{oa})_t$ is the true value of R_{oa}, $\langle R_{oa}|\ldots\rangle$ is its estimate *via* (5) and S is a suitable scale factor. We observe the following.

(*a*) The high values of R_A for the SAD case mainly arise from the twofold ambiguity of the estimates. R_A does not vary with the wavelength because we used calculated data without errors.

(*b*) For $n = 2$, the cases in which f_1'' and f_2'', as well as $\Delta f_1'$ and $\Delta f_2'$, are sufficiently different show better estimates.

(*c*) The use of three wavelengths provides (on average) better results than the use of two, but the accuracy depends on the selected wavelengths (the case in which $\Delta f'$ and f'' are

nearly equal for two wavelengths does not provide good results).

(*d*) The case $n = 4$ does not always provide better estimates than the cases with $n = 3$.

It may be wondered whether our probabilistic approach is able to automatically recognize the different amount of information contained in the different sets of wavelengths and to provide consequent reliabilities of the estimates. A good probabilistic approach should assign larger variances to the R_{oa} estimates for unfavourable wavelength sets. To check this opportunity for each combination of wavelengths quoted in Table 2, we have computed (columns 3 and 5) the average value of the estimates $\langle R_{oa}|\ldots\rangle/\sigma_{R_{oa}}$ as given by (6); we denote such values as $(R_{oa}/\sigma)_{\rm av}$. It is evident [see the high correlation between R_A and $(R_{oa}/\sigma)_{\rm av}$] that our probabilistic approach correctly assigns small variances to the R_{oa} estimates corresponding to the most informative sets of wavelengths. This suggests implementation of the following practical procedure: derive the R_{oa} estimates for each combination of wavelengths, and then select, for location of the anomalous scatterers, the one with the largest value of $(R_{oa}/\sigma)_{\rm av}$.

As an effect of the experimental errors, that which is true for the calculated data is unfortunately not true for the experimental data [see the corresponding values of $(R_{oa}/\sigma)_{\rm av}$ in Table 2]. In particular, the suggestion of selecting the wavelength combination having the largest value of $(R_{oa}/\sigma)_{\rm av}$ is invalidated.

Table 2
ApD and cyanase data.

Values of R_A and of $(R_{oa}/\sigma)_{\rm av}$ for the various combinations of wavelengths (WL). The $\Delta f'$ and f'' values employed for the calculated data at the wavelengths used are as follows: ApD, (−1.805, 0.646), (−8.582, 3.843), (−7.663, 3.841), (−2.618, 3.578); cyanase, (−2.112, 0.595), (−9.643, 0.499), (−8.582, 3.843), (−2.618, 3.578).

	ApD (calc. data)		Cyanase (calc. data)		ApD (exp. data)		Cyanase (exp. data)	
WL	R_A	$(R_{oa}/\sigma)_{\rm av}$	R_A	$(R_{oa}/\sigma)_{\rm av}$	R_A	$(R_{oa}/\sigma)_{\rm av}$	R_A	$(R_{oa}/\sigma)_{\rm av}$
1	32.3	4.9	33.5	4.2	47.2	2.8	55.8	3.4
2	32.1	4.3	33.1	4.2	41.6	3.8	40.8	2.5
3	32.1	4.3	33.3	4.0	40.4	2.9	39.8	2.6
4	32.3	4.3	33.7	4.2	41.4	3.7	39.6	3.4
1-2	13.6	11.3	30.3	4.4	41.4	3.2	39.5	2.9
1-3	6.8	203.0	13.3	8.1	34.9	6.9	41.3	2.8
1-4	31.1	4.2	33.7	4.3	42.0	3.9	47.4	2.8
2-3	32.1	4.5	33.3	4.1	39.7	3.1	39.7	2.8
2-4	31.9	5.2	21.2	5.8	38.4	4.4	43.1	2.8
3-4	33.1	5.4	15.2	8.8	37.5	5.4	44.5	2.9
1-2-3	7.8	264.3	20.6	6.5	36.3	6.1	39.8	3.0
1-2-4	14.8	21.7	25.7	4.8	39.1	3.9	42.6	3.5
1-3-4	22.7	7.7	4.9	18.9	35.0	6.3	43.9	3.5
2-3-4	32.6	5.3	3.6	19.1	38.8	9.1	44.3	2.8
1-2-3-4	19.9	9.3	15.9	7.8	35.8	6.8	41.6	3.5

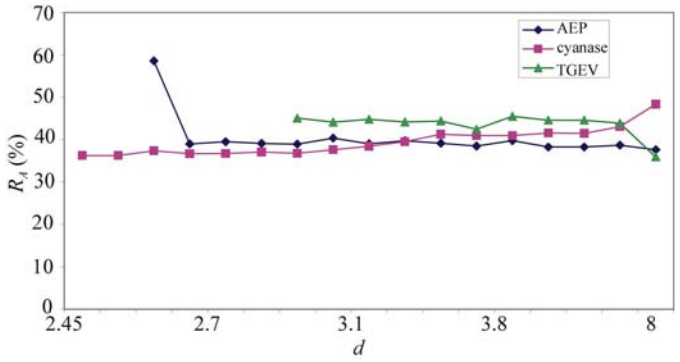

Figure 1
The R_A value *versus* the resolution for three test structures when the experimental data from three wavelengths are used simultaneously.

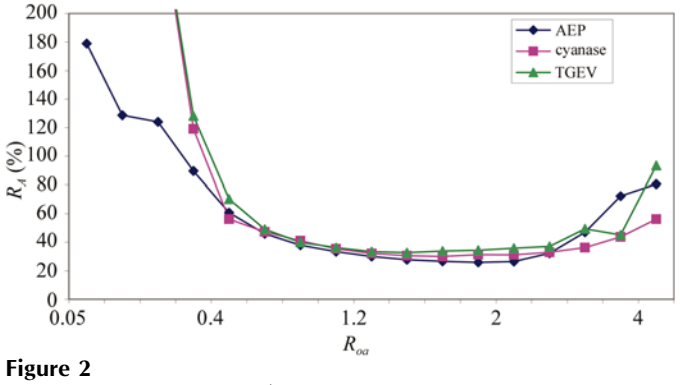

Figure 2
The R_A value *versus* $\langle R_{oa}|\ldots\rangle$ for three test structures when the experimental data from three wavelengths (the same as in Fig. 1) are used simultaneously.

(iii) *Is the error of the estimate dependent on the resolution?* We have no practical evidence of such behaviour: R_A is practically constant with resolution, as may be deduced from Fig. 1, where, for the experimental data of AEP, cyanase and TGEV, the R_A values are plotted against the resolution.

(iv) *Is the error of the estimate dependent on* $\langle R_{oa}|\ldots\rangle$*?* We have clear evidence for this trend. In Fig. 2 we show, for the experimental data of AEP, cyanase and TGEV, the plot of R_A against $\langle R_{oa}|\ldots\rangle$. Quite wrong estimates are frequent for small values of $\langle R_{oa}|\ldots\rangle$; the best estimates are attained for medium values of $\langle R_{oa}|\ldots\rangle$, while a loss of accuracy may be noted for the largest $\langle R_{oa}|\ldots\rangle$ values. This last behaviour, even if it concerns a limited number of estimates, is not ideal for the application of direct methods, the success of which is based on the accuracy of the largest structure-factor moduli.

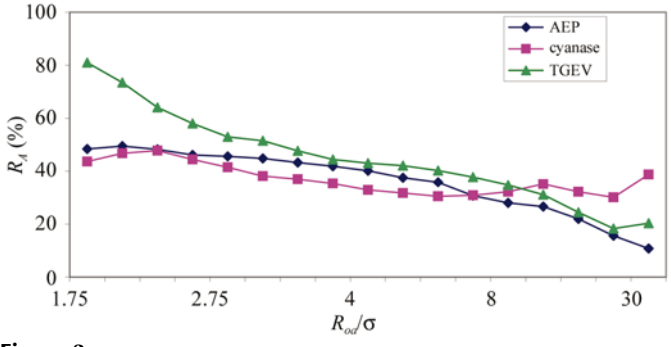

Figure 3
The R_A value *versus* $\langle R_{oa}|\ldots\rangle/\sigma_{R_{oa}}$ for three test structures when the experimental data from three wavelengths (the same as in Figs. 1 and 2) are used simultaneously.

A more useful result is obtained if (6) is used. It has been shown in paper I that high values of $\langle R_{oa}|\dots\rangle$ do not necessarily correlate with sharp distributions (2): *e.g.* large values of $\langle R_{oa}|\dots\rangle$ may be coupled with small or with large values of $\sigma_{R_{oa}}$, and medium-size values of $\langle R_{oa}|\dots\rangle$ may show small or large $\sigma_{R_{oa}}$ values. If this result is extended to the n-wavelength case, it may be guessed that smaller values of R_A should be obtained for subsets of reflections characterized by small values of $\langle R_{oa}|\dots\rangle/\sigma_{R_{oa}}$. This expectation is confirmed in Fig. 3, where we plot R_A against $\langle R_{oa}|\dots\rangle/\sigma_{R_{oa}}$. The reader can usefully compare the relatively high R_A values of the reflections with the largest values of $\langle R_{oa}|\dots\rangle$ with the relatively small R_A values of the reflections with the largest values of $\langle R_{oa}|\dots\rangle/\sigma_{R_{oa}}$. Accordingly, we decided to select the reflections (for direct-methods applications) in order of $\langle R_{oa}|\dots\rangle/\sigma_{R_{oa}}$ rather than in order of $\langle R_{oa}|\dots\rangle$.

5. The phasing procedure

In accordance with the results established in §4, the following procedure is used to find the positions of the anomalous scatterers.

Step 1. The program reads and stores the sets $S_j, j = 1, \dots, n$, of the observed magnitudes (say $|F^+|$, $|F^-|$) for all n wavelengths.

Step 2. Each S_j is placed on an absolute scale by the Wilson method and the corresponding overall thermal factor is estimated.

Step 3. Matrix **K** is calculated and (5) and (6) are applied to obtain the values $\langle R_{oa}|\dots\rangle$ and $\langle R_{oa}|\dots\rangle/\sigma_{R_{oa}}$. To eliminate *ab initio* the less informative cases, we omit from the next calculations the reflections for which, for at least one wavelength, only one of $|F^+|$ and $|F^-|$ is measured.

Step 4. The three-phase invariants involving the reflections with the highest $\langle R_{oa}|\dots\rangle/\sigma_{R_{oa}}$ values are evaluated. We experimentally checked the relative usefulness of the Cochran (1955) formula and of the P_{10} formula (Cascarano *et al.*, 1984). We found the second formula more useful; therefore, the P_{10} formula is the default choice of the program.

Step 5. The tangent formula is used (random starting approach). Figures of merit (see below) are used to reduce the number of trial solutions to admit to next steps.

Step 6. The direct-space refinement techniques used in *SIR*2002 (Burla *et al.*, 2002) are used to extend the phase information (gained in step 5) to a larger set of reflections: only 30% of the reflections with the smallest values of $\langle R_{oa}|\dots\rangle$ remain unphased. The final calculations of this step are constituted by automatic cycles of least-squares refinement, which aim at refining the substructure model provided by the trial solutions.

Step 7. Suitable figures of merit (see below) are used to recognize the correct substructure models.

Let us now examine the most critical points of the procedure.

(i) The FOMs used in step 5. If the $\langle R_{oa}|\dots\rangle$ values were very accurate, the classical figures of merit used in the multi-solution procedures (Germain *et al.*, 1970; Cascarano *et al.*, 1987, 1992) should easily recognize the true solutions among the various trials (the anomalous effects reduce the structure complexity from N, the number of atoms in the unit cell, to a, the number of the anomalous scatterers). Unfortunately, the R_{oa} estimates present a non-negligible variance, which is particularly high for small values of $\langle R_{oa}|\dots\rangle$. Therefore, the powerful figures of merit (PSCOMB and CPHASE) based on the psi0 triplets, negative triplets and negative quartets cannot be applied. In our procedure, we use only two figures of merit, MABS and ALFCOMB (see Cascarano *et al.*, 1992): the combined figure CFOM is used to eliminate from the next calculations half of the trial solutions.

(ii) The FOMs used in step 7. Two figures of merit, RAT and RES, are calculated for those trial solutions which overcome the figure of merit filter defined in step 5.

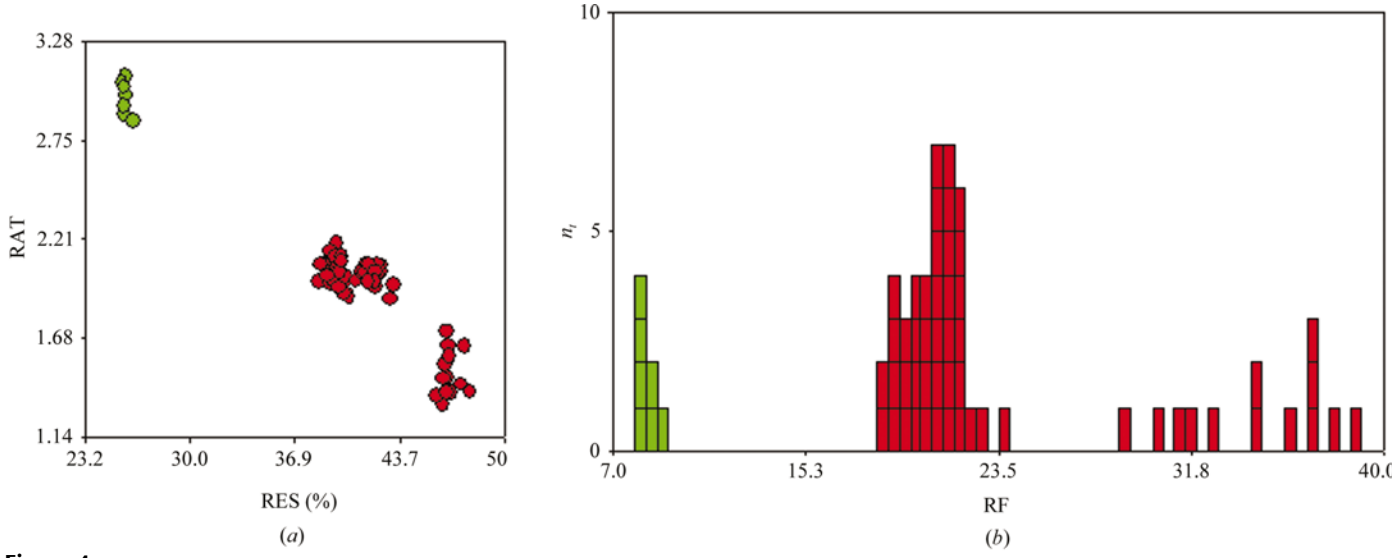

Figure 4
(*a*) Scatter plot of RAT *versus* RES for Tm0665 showing clusters of correct solutions (green) and wrong solutions (red). (*b*) The corresponding RF histogram: the number of trials (n_t) *versus* RF.

(*a*) The first figure, RAT, is similar to that used by *SIR*2002 (Burla *et al.*, 2002),

$$\mathrm{RAT} = \mathrm{CC}/\langle R_{\rm cal}^2 \rangle,$$

where

$$CC = \frac{\left(\left\langle \langle R_{oa}|...\rangle^2 \cdot w^2 \right\rangle - \left\langle \langle R_{oa}|...\rangle^2 \right\rangle \cdot \langle w^2 \rangle\right)}{\left(\left\langle \langle R_{oa}|...\rangle^4 \right\rangle - \left\langle \langle R_{oa}|...\rangle^2 \right\rangle^2\right)^{1/2} \cdot (\langle w^4 \rangle - \langle w^2 \rangle^2)^{1/2}},$$

where $w = D_1(\langle R_{oa}|\ldots\rangle R_{\rm cal}/2)$ is a SIM-like weight, $D_1(x) = I_1(x)/I_0(x)$ is the ratio of the modified Bessel functions of order one and zero, respectively, $\langle R_{oa}|\ldots\rangle$ plays the role of the observed modulus and $R_{\rm cal}$ is the corresponding modulus calculated from the substructure model available at that moment.

The numerator of RAT involves the weights and the reflections actively used in the phasing process (about 70% of the measured reflections) and is expected to be maximum for the correct solution (this occurs if the phases of the largest 'observed' moduli are defined with large weights). The denominator of RAT is calculated for the reflections not involved in the phasing procedure (about the 30% of the total, as stated at step 6) and it is expected to be minimum for the correct solution.

Accordingly, RAT should be maximum for the most promising trial solutions.

(*b*)

$$\mathrm{RES} = \frac{\sum \left| |F_{oa}|_{\rm calc} - \langle F_{oa}|...\rangle \right|}{\sum \langle F_{oa}|...\rangle},$$

where $|F_{oa}|_{\rm calc}$ is the value of F_{oa} calculated from the model at the end of the least-squares refinement process, $\langle F_{oa}|\ldots\rangle$ is obtained by multiplying $\langle R_{oa}|\ldots\rangle$ by the scale factor and the overall thermal factor defined by the Wilson plot. It may be expected that large values of RAT and small values of RES will select the correct solutions. We show, in Figs. 4(*a*) and 5(*a*) for Tm0665 and AEP, respectively, the distribution of the solutions in the plane (RAT, RES). The clustering of the correct solutions is more evident for Tm0665, but in both the cases we were able to select the correct solutions among the various trials.

We have therefore decided to use the ratio

$$\mathrm{RF} = (\mathrm{RES} \times 100)/\mathrm{RAT}$$

as an automatic tool to separate the correct from the wrong solutions; RF is expected to be a minimum for the good solutions. In Figs. 4(*b*) and 5(*b*) we show the RF histograms for Tm0665 and AEP, respectively.

RF is able to automatically select the good solutions for all our test structures.

6. The substructure models

The procedure described in §5 has been applied to all the test structures quoted in Table 1. The robustness of the method allowed us to limit the number of trials (in the tangent formula step) to a maximum of 60 for all the applications. The results (experimental diffraction data) are shown in Tables 3 and 4.

Let us first examine the SAD case (eight test structures). For CauFd (see Table 3) only one solution is found; in all the other cases the density of good solutions is higher than 1/60. The completeness of the models is also satisfactory (see Table 3): all or nearly all the anomalous scatterers are correctly located.

The MAD case (eight structures) allows different attempts according to the specific wavelengths used for obtaining the estimates provided by (5) and (6). In the first part of Table 3 we quote, for a selected wavelength combination, the number and the quality of the correct solutions. A complete overview of the results for each structure and for each wavelength

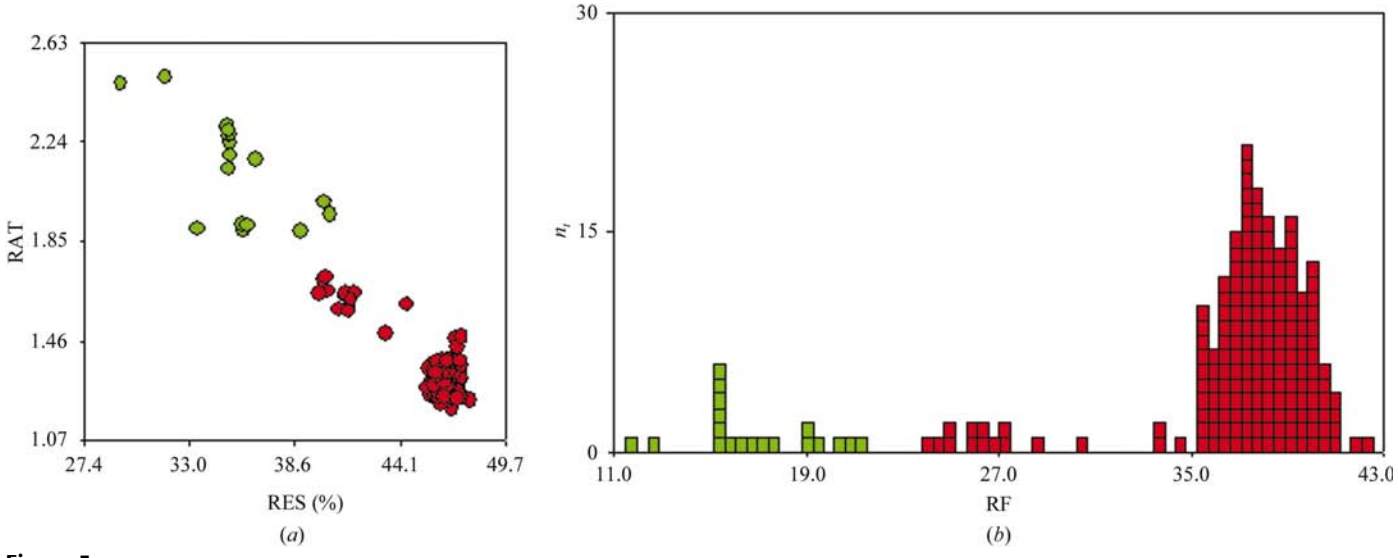

Figure 5
(*a*) Scatter plot of RAT *versus* RES for AEP showing clusters of correct solutions (green) and wrong solutions (red). (*b*) The corresponding RF histogram: the number of trials (n_i) *versus* RF.

Table 3
Results for the test structures.

WL describes the wavelength combination, n_{sol} is the number of solutions and n_{af} is the number of anomalous scatterers located.

Protein code	WL	n_{sol}	n_{af}
ApD	1-2-3-4	6	3/3
JIA	1-2-3-4	5	8/8
KPR	1-2-3	6	8/8
PSCP	1-2-3	25	12/13
Cyanase	1-2-3	40	40/40
Tm0665	1-2-3	12	44/45
TGEV	1-2-3-4	10	56/60
AEP	1-2-3	9	66/66
Glucose isomerase	1	2	1/1
2Zn insulin	1	53	2/2
Ca-Subtilisin	1	3	3/3
CauFd	1	1	8/8
DNA	1	4	10/10
Lysozyme	1	2	17/17
CUTA1	1	2	16/18
APT	1	6	16/22

Table 4
Number of correct solutions over 60 trials for the MAD test structures.

WL describes the wavelength combination; dashes denote forbidden combinations (for the three-wavelength data).

WL	ApD	JIA	KPR	PSCP	Cyanase	Tm0665	TGEV	AEP
1	0	0	8	11	0	0	0	6
2	1	0	2	7	55	0	3	5
3	2	1	10	22	60	6	0	4
4	3	4	—	—	2	—	0	—
1-2	3	0	5	0	19	0	0	5
1-3	4	2	8	18	0	8	0	6
1-4	2	4	—	—	0	—	0	—
2-3	2	6	10	10	59	3	0	7
2-4	7	2	—	—	0	—	8	—
3-4	7	2	—	—	0	—	5	—
1-2-3	3	3	6	25	40	12	0	9
1-2-4	4	3	—	—	0	—	4	—
1-3-4	3	6	—	—	0	—	2	—
2-3-4	3	7	—	—	0	—	9	—
1-2-3-4	6	5	—	—	0	—	10	—

combination is shown in Table 4. In the two tables WL describes the wavelength combination, n_{sol} is the number of solutions (over 60 trials): for brevity, n_{af}, the number of anomalous scatterers automatically located, is only quoted in Table 3 (when the correct solution is found, its quality is nearly independent of the wavelength combination). Table 4 shows that the various wavelength combinations are not equally informative: a lot of solutions may be found for some of them, while no correct solution may be identified for others. Thus, the capacity of working with any wavelength combination is a reserve of power which cannot be overlooked, particularly in difficult cases. Our method proved able to find the correct solution for all the test structures.

7. Conclusions

The probabilistic theory started in paper I for obtaining, from two-wavelength data, the structure-factor moduli of the anomalous scatterer substructure has been generalized to SAD and MAD cases. A procedure has been devised which automatically performs all the steps necessary to locate the anomalous scatterers. The success of such a procedure depends on several ingredients: the specific combination of wavelengths, the efficiency of our probabilistic method of estimating R_{oa}, the number of trials explored in the direct-methods procedure and the efficiency of the direct-space phase-refinement process. Applications to a large set of test structures proved the efficiency and robustness of our method.

Table 5
The matrix **K** (boxed) for $n = 2$.

The two rows (and columns) outside the box indicate the variables with respect to which the elements of the matrix are calculated and their order number.

		A_{oa}	A_1^+	A_2^+	A_1^-	A_2^-	B_{oa}	B_1^+	B_2^+	B_1^-	B_2^-
		1	2	3	4	5	6	7	8	9	10
A_{oa}	1	—	$\langle A_{oa}A_1^+\rangle$	—	—	—	—	—	—	—	—
A_1^+	2	—	—	—	—	—	—	$\langle A_1^+B_1^+\rangle$	—	—	—
A_2^+	3	—	—	—	—	$\langle A_2^+A_2^-\rangle$	—	—	—	—	—
A_1^-	4	—	—	—	—	—	—	—	—	—	—
A_2^-	5	—	—	—	—	—	—	—	—	—	—
B_{oa}	6	—	—	—	—	—	—	—	—	—	—
B_1^+	7	—	—	—	—	—	—	—	—	—	—
B_2^+	8	—	—	—	—	—	—	—	—	—	—
B_1^-	9	—	—	—	—	—	—	—	—	—	—
B_2^-	10	—	—	—	—	—	—	—	—	—	—

Table 6
Some additional columns of the matrix **K** for $n = 3$.

		A_{oa}	A_1^+	A_2^+	A_3^+	A_1^-	A_2^-	A_3^-	B_{oa}	B_1^+	B_2^+	B_3^+	B_1^-	B_2^-	B_3^-
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
A_{oa}	1	—	—	—	k_{14}	—	—	k_{17}	—	—	—	$k_{1,11}$	—	—	$k_{1,14}$
A_1^+	2	—	—	—	k_{24}	—	—	k_{27}	—	—	—	$k_{2,11}$	—	—	$k_{2,14}$
A_2^+	3	—	—	—	k_{34}	—	—	k_{37}	—	—	—	$k_{3,11}$	—	—	$k_{3,14}$
—		—	—	—	—	—	—	—	—	—	—	—	—	—	—
—		—	—	—	—	—	—	—	—	—	—	—	—	—	—
B_3^-	14	—	—	—	$k_{14,\,4}$	—	—	$k_{14,7}$	—	—	—	$k_{14,11}$	—	—	$k_{14,14}$

APPENDIX *A*

Let us consider in Table 5 the geometry of the **K** matrix for the case $n = 2$: inside the box some elements of the matrix are emphasized (*i.e.* k_{12}, k_{27}, k_{35}), outside the box, in two rows and in two columns, the reference variables and their order number [as they appear in the distribution (1)] are shown.

The emphasized **K** elements have the following expressions (see paper I):

$$\begin{aligned} k_{12} &= \langle A_{oa}A_1^+\rangle = S_9/(\Sigma_{oa}\Sigma_{N1})^{1/2} \\ &= (\Sigma_a f_j^o f'_{j1})/[(\Sigma_a f_j^{o2})\Sigma_N(f'^2_{j1} + f''^2_{j1})]^{1/2}, \\ k_{27} &= \langle A_1^+B_1^+\rangle = 0, \\ k_{35} &= \langle A_2^+A_2^-\rangle = (\Sigma_o + S_2)/\Sigma_{N2} \\ &= [(\Sigma_{Na}f_j^{o2}) + \Sigma_a(f'^2_{j2} - f''^2_{j2})]/\Sigma_N(f'^2_{j2} + f''^2_{j2}). \end{aligned}$$

If we want to construct the matrix **K** for the case $n = 3$, we have to introduce into the matrix the additional columns shown in Table 6 (and the corresponding rows). Such rows and columns represent the covariance terms between the third wavelength

and the first two. It is easily understood that for the case $n = 3$ the elements of the columns k_{j4}, k_{j7}, k_{j11}, k_{j14} (all relative to the third wavelength) can be calculated by analogy with the terms k_{j3}, k_{j6}, k_{j10} and k_{j13} of the same matrix (all relative to the second wavelength) or, equivalently, by analogy with the terms k_{j2}, k_{j5}, k_{j9} and k_{j12} (relative to the first wavelength). For example, for $n = 3$,

$$k_{14} = \langle A_{oa} A_3^+ \rangle = (\Sigma_a f_j^o f'_{j3})/[(\Sigma_a f_j^{o2})\Sigma_N (f'^2_{j3} + f''^2_{j3})]^{1/2},$$
$$k_{13} = \langle A_{oa} A_2^+ \rangle = (\Sigma_a f_j^o f'_{j2})/[(\Sigma_a f_j^{o2})\Sigma_N (f'^2_{j2} + f''^2_{j2})]^{1/2}$$

and k_{12} (see the first algebraic expression in this section) can be obtained from the others by changing the wavelength-dependent values f' and f''.

In this recursive way, the matrix **K** may be constructed for any value of n.

As soon as the matrix **K** has been obtained, the various terms of the quadratic form $\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T}$ [see (2)] can be derived by associating each λ_{ij} element with two elements of the vector **T**. For example, λ_{ij} has to be associated with the two reference variables with order numbers i and j, as has been performed for (1) in the two-wavelength case.

We thank Dr Dritan Siliqi for enlightening discussions. We thank also Drs Z. Dauter, C. M. Weeks, S. Mangani, V. Calderone, R. Hilgenfeld, D. Matak, M. A. Walsh and A. Gonzalez for having kindly provided us with the experimental diffraction data.

References

Anand, K., Palm, G. J., Mesters, J. R., Siddell, S. G., Ziebuhr, J. & Hilgenfeld, R. (2002). *EMBO J.* **21**, 3213–3224.
Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (2001). *J. Appl. Cryst.* **34**, 523–526.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2003). In the press.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2002). *Acta Cryst.* D**58**, 928–935.
Calderone, V., Benvenuti, M., Viezzoli, M. S., Bertini, I. & Mangani, S. (2002). *Z. Kristallogr.* **217**, 629–635.
Cascarano, G. L., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* A**40**, 278–283.
Cascarano, G. L., Giacovazzo, C. & Guagliardi, A. (1992). *Acta Cryst.* A**48**, 859–865.
Cascarano, G. L., Giacovazzo, C. & Viterbo, D. (1987). *Acta Cryst.* A**43**, 22–29.
Chen, C. C. H., Kim, A., Zhang, H., Howard, A. J., Sheldrick, G. M., Dunaway-Mariano, D. & Herzberg, O. (2000). Abstr. Am. Crystallogr. Assoc. Meet., Abstract 02.06.03.
Cochran, W. (1955). *Acta Cryst.* **8**, 473.
Dauter, Z., Dauter, M. & Dodson, E. (2002). *Acta Cryst.* D**58**, 494–506.
Dauter, Z., Li, M. & Wlodawer, A. (2001). *Acta Cryst.* D**57**, 239–249.
Foadi, J., Woolfson, M. M., Dodson, E., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.
Germain, G., Main, P. & Woolfson, M. M. (1970). *Acta Cryst.* B**26**, 274–285.
Giacovazzo, C. (1983). *Acta Cryst.* A**39**, 585–592.
Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). *Acta Cryst.* D**55**, 1568–1577.
Hauptman, H. A. (1982). *Acta Cryst.* A**38**, 632–641.
Hendrickson, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11–21.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Howell, P. L., Blessing, R. H., Smith, G. D. & Weeks, C. M. (2000). *Acta Cryst.* D**56**, 604–617.
Karle, J. (1980). *Int. J. Quantum Chem. Quantum Biol. Symp.* **7**, 357–367.
Lesley, S. A. *et al.* (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
Li, J., Derewenda, U., Dauter, Z., Smith, S. & Derewenda, Z. S. (2000). *Nature Struct. Biol.* **7**, 555–559.
Matak-Vinkovic, D., Vinkovic, M., Saldanha, S. A., Ashurst, J. A., Von Delft, F., Inoue, T., Miguel, R. N., Smith, A. G., Blundell, T. L. & Abell, C. (2001). *Biochemistry*, **40**, 14493.
Miller, R., Gallo, S. M., Khala, M. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.
Pähler, A., Smith, J. L. & Hendrickson, W. A. (1990). *Acta Cryst.* A**46**, 537–540.
Schneider, R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Sheldrick, G. M. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 131–141. Dordrecht: Kluwer Academic Publishers.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. (1998). *Acta Cryst.* D**54**, 799–804.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 11–16.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Walsh, M. A., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* D**55**, 1168–1173.
Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (2000). *Structure*, **8**, 505–514.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.