Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

MAD phasing: choosing the most informative wavelength combination

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

[a]Dipartimento di Scienze della Terra, Piazza Universita', 06100 Perugia, Italy, [b]Istituto di Cristallografia, CNR, Via G. Amendola 122/O, 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 3 May 2004
Accepted 29 June 2004

Two algorithms are described for limiting data resolution and for predicting the most informative wavelength combinations in MAD techniques. Both have been successfully tested using experimental data from a large set of test structures.

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.
ε: statistical Wilson coefficient.
$f_j = f_j^0 + \Delta f_j + if_j'' = f_j' + if_j''$: 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 the atoms in the unit cell.
$\Sigma_o = \sum_{j=1}^{na}(f_j^o)^2$. The summation is extended to all the 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.
$F^{\pm} = |F^{\pm}|\exp(i\varphi^{\pm}) = F_{\pm\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(\pm 2\pi i \mathbf{h}\mathbf{r}_j)$.
$E^{+} = F^{+}/(\varepsilon\Sigma_N)^{1/2} = R\exp(i\varphi^{+}) = A^{+} + iB^{+}$.
$E^{-} = F^{-}/(\varepsilon\Sigma_N)^{1/2} = G\exp(i\varphi^{-}) = A^{-} + iB^{-}$.
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.
$F_{oa} = |F_{oa}|\exp(i\varphi_{oa}) = \sum_{j=1}^{a} f_j^o \exp(2\pi i \mathbf{h}\mathbf{r}_j)$.
$E_{oa} = F_{oa}/(\varepsilon\Sigma_{oa})^{1/2} = R_{oa}\exp(i\varphi_{oa}) = A_{oa} + iB_{oa}$.
$\Delta_{\text{ano}} = |F^{+}| - |F^{-}|$.

2. Introduction

The traditional SAD or MAD procedure is substantially a two-step process: the anomalous scatterer substructure is first determined and refined (Karle, 1980; Hendrickson, 1985; Pähler *et al.*, 1990; Terwilliger, 1994; Sheldrick *et al.*, 1993; Sheldrick, 1998; Terwilliger & Berendzen, 1999; Grosse-Kunstleve & Brunger, 1999) and then protein phases are assigned. The practice of introducing Se atoms into a protein as selenomethionines encouraged the use (to define the anomalous scatterer substructure) of the last generation of direct-methods programs (Howell *et al.*, 2000; Schneider & Sheldrick, 2002; Burla, Camalli *et al.*, 2003; Foadi *et al.*, 2000).

A new approach has been suggested in two recent papers (Burla *et al.*, 2002; Burla, Carrozzini, Cascarano *et al.*, 2003): the estimates of the amplitudes of the structure factors of the anomalously scattering substructure are derived, *via* the rigorous method of the joint probability distribution functions, from the experimental diffraction moduli relative to n wavelengths. To do that, first the joint distribution

$$\begin{aligned} P_n &= P(A_{oa}, A_1^{+}, A_2^{+}, \ldots, A_n^{+}, A_1^{-}, A_2^{-}, \ldots, \\ &\quad A_n^{-}, B_{oa}, B_1^{+}, B_2^{+}, \ldots, B_n^{+}, B_1^{-}, B_2^{-}, \ldots, B_n^{-}) \\ &= \pi^{-(2n+1)}(\det \mathbf{K})^{-1/2}\exp(-\tfrac{1}{2}\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T}) \end{aligned} \quad (1)$$

is calculated, 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^{-}. Then, the conditional distribution

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

is derived, from which

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

is obtained, where

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

The standard deviation of the estimate is also calculated:

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

from which

$$\frac{\langle R_{oa}|\ldots\rangle}{\sigma_{R_{oa}}} = \left[\frac{(\pi/4) + (X^2)/\lambda_{11}}{1 - (\pi/4)}\right]^{1/2}. \quad (3)$$

The advantage of the above approach is that the estimates can simultaneously exploit both the anomalous and the dispersive differences. The computing procedure proposed by Burla, Carrozzini, Cascarano *et al.* (2003) is as follows.

 DOI: 10.1107/S0907444904015835

Table 1
Set of test structures.

n_{wl} is the number of wavelengths used in the experiment; *na* is the number of symmetry-independent anomalous scatterers.

Protein code	Space group	n_{wl}	Anomalous scatterer	*na*	Resolution (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)
IDI	$P4_12_12$	2	Se	8	2.4	Bonanno *et al.* (2001)
MDD	$P2_12_12$	3	Se	9	2.3	Bonanno *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)
KSM	$P2_1$	3	Se	160	2.9	Von Delft *et al.* (2003)

(i) The sets $S_j, j = 1, \ldots, n$, of the observed magnitudes (say $|F^+|$, $|F^-|$) are stored for all n wavelengths.

(ii) The Wilson method is applied to put the sets S_j onto absolute scales.

(iii) Equations (2) and (3) are applied to obtain the values $\langle R_{oa}|\ldots.\rangle$ and $\langle R_{oa}|\ldots\rangle/\sigma_{R_{oa}}$.

(iv) The triplet invariants involving the reflections with the highest $\langle R_{oa}|\ldots\rangle/\sigma_{R_{oa}}$ values are evaluated and the the tangent formula is applied *via* a random starting approach.

(v) The direct-space refinement techniques of *SIR*2002 (Burla, Camalli *et al.*, 2003) are used to extend the phase information to a larger set of reflections: only 30% of the reflections with the smallest values of $\langle R_{oa}|\ldots\rangle$ remain unphased. Automatic cycles of least-squares refinement improve the substructure model provided by the trial solutions.

(vi) Suitable figures of merit are used to recognize the correct substructure models.

The application of the above procedure to several MAD cases showed that (i) the capacity to use any wavelength combination to find the anomalous scatterer substructure is a reserve of power which cannot be overlooked, particularly in difficult cases, and (ii) the various wavelength combinations are not equally informative: a lot of correct solutions can be found for some of them, while no correct solution may be identified for others.

Unfortunately, the above procedure is not accompanied by any criterion able to predict the most informative wavelength combinations. Thus, to safely determine the substructure when $n = 4$, the user should explore four one-wavelength combinations, six two-wavelength combinations, four three-wavelength combinations and one four-wavelength combination and rely on proper figures of merit in order to identify the correct solutions. Providing a criterion for predicting the most informative combinations is the aim of this paper, which will also indicate a filtering criterion that can make the method more robust and efficient. Both the criteria suggested here are based on a basic assumption: good experimental multi-wavelength data should show high correlation between the various Δ_{ano} values. Such a premise has been already recognized by Schneider & Sheldrick (2002).

3. Filtering: the algorithm

It is quite frequent that the correlation between $\Delta_{ano}(i)$ and $\Delta_{ano}(j)$ (where i and j denote the ith and jth wavelengths) decreases with $\sin\theta/\lambda$. Overly large decrements of the correlation indicate lower quality of the high-resolution diffraction data. Schneider & Sheldrick (2002) advise choosing the high-energy remote wavelength as a reference for calculating the correlation coefficients and truncating the data where the correlation coefficient falls below about 25–30%. The algorithm described here tries to choose for each wavelength combination a threshold for the data resolution in order to eliminate the experimental multiwavelength data with low correlation coefficients between the various Δ_{ano} values.

The observed $\sin\theta/\lambda$ range has been divided into p_{tot} (depending on the data resolution) intervals.

(i) For each interval p and for each pair of wavelengths (i, j) the values $\mathrm{COR}(i, j; p)$ have been calculated, where $\mathrm{COR}(i, j; p)$ is the correlation factor between $\Delta_{ano}(i)$ and $\Delta_{ano}(j)$ for the pth $\sin\theta/\lambda$ interval.

(ii) For each interval p and for each triplet of wavelengths (i, j, k) the values

$$\mathrm{COR}(i, j, k; p) = \mathrm{WT3}[\mathrm{COR}(i, j; p) + \mathrm{COR}(i, k; p) + \mathrm{COR}(j, k; p)]/3$$

are calculated where WT3 = 1.2.

(iii) For each interval p and for each quadruplet (i, j, k, l) the values

$$\mathrm{COR}(i, j, k, l; p) = \mathrm{WT4}[\mathrm{COR}(i, j; p) + \mathrm{COR}(i, k; p) + \mathrm{COR}(i, l; p) + \mathrm{COR}(j, k; p) + \mathrm{COR}(j, l; p) + \mathrm{COR}(k, l; p)]/6$$

are calculated, where WT4 = 1.25.

For five wavelengths analogous values of COR are calculated. For each λ-combination

(i) the average value of the first three $\sin\theta/\lambda$ intervals

$$\mathrm{AMED}(\lambda\text{-combination}) = \langle \mathrm{COR}(\lambda\text{-combination}; p = 1 \text{ to } 3)\rangle$$

is calculated,

(ii) the value of last is determined according to the condition

$$\mathrm{COR}(\lambda\text{-combination; last}) < \mathrm{PERC} \times \mathrm{AMED}(\lambda\text{-combination}),$$

where PERC = 0.25 (scarcely populated intervals are not taken into account) and

(iii) data starting from the last + 1 interval are omitted from the next calculations.

At the end of the above procedure a threshold on the data resolution (thrres) is obtained for each λ-combination.

4. The choice of the most promising λ-combinations: the algorithm

The following rules are applied to choose the most effective λ-combination.

(i) For each λ-pair the value $\mathrm{CC}(i, j)$, the correlation factor between $\Delta_{ano}(i)$ and $\Delta_{ano}(j)$ (i, j define the two wavelengths), is calculated for all the data selected by the algorithm described in the preceding paragraph.

(ii) The value of $\mathrm{RANK}(i) = \langle \mathrm{CC}(i, j)\rangle_{j\neq i}$ is calculated for each ith wavelength.

(iii) For each λ-pair the value $\mathrm{RANK}(i, j) = \mathrm{WT2} \times \mathrm{CC}(i, j)$ is set, where WT2 = 1.1.

(iv) For each λ-triplet

$$\mathrm{RANK}(i, j, k) = \mathrm{WT3} \times [\mathrm{CC}(i, j) + \mathrm{CC}(i, k) + \mathrm{CC}(j, k)]/3$$

is calculated, where WT3 = 1.1.

(v) For each λ-quadruplet the value

$$\mathrm{RANK}(i,j, k, l) = \mathrm{WT4} \times [\mathrm{CC}(i, j) + \mathrm{CC}(i, k) + \ldots + \mathrm{CC}(k, l)]/6$$

is calculated, where WT4 = 1.3.

(vi) An analogous calculation for the eventual combination of five λs.

Table 2
For each test structure we show the possible wavelength combinations (λ_{comb}) and the corresponding RANK values.

n_{sol} is the number of attained solutions over 60 trials (300 trials were used only for KSM), *nf* is the number of anomalous scatterers located by the procedure and thrres is the data resolution chosen by the program (a bar indicates that thrres = Resol).

ApD	λ comb	2-3-4	2-3	3-4	2-4	3	1-2-3-4	2	4	1-2-3	1-3-4	1-2-4	1-3	1	1-4	1-2
	RANK	85.6	80.3	77.0	68.9	60.6	59.4	55.0	54.9	45.6	45.5	39.1	34.5	19.6	18.7	15.7
	$_{nsol}$/*nf*	8/3	1/3	5/3	5/3	6/3	2/3	8/3	2/3	7/3	4/3	4/3	7/3	0	1/3	2/3
	thrres	—	—	—	—	2.5	—	—	2.5	—	—	—	2.5	2.5	2.5	—
JIA	λ $_{comb}$	2-3-4	3-4	2-3	2-4	1-2-3-4	4	3	2	1-3-4	1-2-4	1-2-3	1-4	1-2	1	1-3
	RANK	66.5	66.3	55.5	54.1	53.6	51.2	47.2	46.0	45.2	43.7	39.2	33.2	28.3	27.1	19.9
	n_{sol}/*nf*	7/8	4/8	3/8	0	5/8	2/8	4/8	1/8	1/8	0	1/8	4/8	1/8	0	3/8
	thrres	2.5	2.5	2.5	2.6	2.6	3.2	2.5	3.0	2.6	2.8	2.6	3.2	3.0	3.2	—
KPR	λ_{comb}	1-2-3	1-2	1	2	1-3	3	2-3								
	RANK	63.6	60.8	58.3	56.1	55.8	53.7	51.5								
	n_{sol}/*nf*	18/8	12/8	14/8	13/8	9/8	4/8	12/8								
	thrres	2.6	2.6	3.0	2.6	3.0	3.0	2.6								
IDI	λ_{comb}	1-2	2	1												
	RANK	59.6	54.2	54.2												
	n_{sol}/*nf*	4/8	7/8	0												
	thrres	—	—	—												
MDD	λ_{comb}	1-2-3	1-2	2	2-3	1	3	1-3								
	RANK	89.5	80.8	80.2	79.6	78.5	77.9	76.2								
	n_{sol}/*nf*	4/9	3/9	9/9	1/9	8/9	4/9	10/9								
	thrres	2.4	—	2.5	2.5	2.5	2.5	2.5								
PSCP	λ_{comb}	1-2-3	1-3	1	1-2	3	2	2-3								
	RANK	81.9	72.7	72.6	72.4	72	71.8	71.2								
	n_{sol}/*nf*	22/12	12/12	8/12	15/12	13/12	13/12	17/12								
	thrres	2.2	2.3	2.3	2.2	2.3	2.2	2.2								
Cyanase	λ_{comb}	2-3	2-3-4	3-4	2-4	1-2-3-4	4	3	2	1-3-4	1-2-4	1-2-3	1-4	1	1-3	1-2
	RANK	92.1	92.0	77.7	73.3	68.6	64.8	63.8	62.2	54	52.2	51.1	43.4	28.8	21.7	21.3
	n_{sol}/*nf*	59/38	50/40	26/40	40/40	59/40	59/40	60/38	60/35	52/40	60/40	60/40	53/40	0	59/40	60/40
	thrres	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Tmo665	λ_{comb}	1-2-3	1-3	3	2-3	1	2	1-2								
	RANK	73.6	66.2	65.6	64.9	64.8	64.1	63.3								
	n_{sol}/*nf*	11/44	9/44	5/43	3/44	0	0	0								
	thrres	2.1	2.1	2.1	2.1	2.1	2.1	—								
TGEV	λ_{comb}	2-3-4	2-4	2-3	3-4	2	1-2-3-4	4	3	1-2-4	1-2-3	1-3-4	1-4	1-2	1	1-3
	RANK	72.8	70.5	64.6	57.3	49.3	47.9	46.9	44.5	36.5	33.7	31.0	13.0	12.9	12.5	11.6
	n_{sol}/*nf*	21/56	31/56	10/55	13/56	2/54	19/56	0	0	21/56	1/51	8/54	0	1/53	0	0
	thrres	—	—	—	—	3.6	—	4.0	4.0	3.0	—	3.2	4.0	3.6	4.0	4.0
AEP	λ_{comb}	1-2-3	1-2	1	2	1-3	3	2-3								
	RANK	101.5	93.6	90.6	90.4	87.6	87.4	87.2								
	n_{sol}/*nf*	19/66	13/66	7/66	14/66	19/66	5/66	11/66								
	thrres	2.6	2.6	2.6	2.6	2.6	2.6	2.6								
KSM	λ_{comb}	1-2-3	1-2	2	1	2-3	3	1-3								
	RANK	58.3	57.5	53.2	52.6	48.8	48.3	47.8								
	n_{sol}/*nf*	6/158	0	0	0	0	0	0								
	thrres	3.6	3.6	3.6	4.0	3.6	4.0	4.0								

All the λ-combinations are ordered according to the values of RANK: the λ-combination with the largest value of RANK is chosen as the most informative. If it does not provide useful results, λ-combinations with smaller values of RANK are explored.

As the amount of information contained in the dispersive differences should never be overlooked, both the algorithms described in §2 and §3 favour (through the weight WT3 and WT4) combinations with more than two wavelengths.

5. Applications

We have implemented the algorithms described in §3 and §4 in a modified version of *SIR*2003-N (Burla, Carrozzini, Caliandro *et al.*, 2003) and we have applied them to the test structures quoted in Table 1. For each structure we give the space group, the number of wavelengths used in the experiment, the atomic species of the anomalous scatterers, the number of symmetry-independent anomalous scatterers and the data resolution. Table 2 shows, for each test structure and for each wavelength combination, the corresponding RANK values, the number of correct solutions attained over 60 trials (300 trials were attempted for KSM only, owing to its size), the values of thrres and of *nf*, where *nf* is the maximum number of substructure atoms found for each wavelength combination. We observe the following.

(i) It is not always the selected combinations that provide the largest solution density, but in all cases high values of RANK select combinations with good density.

(ii) For the same test structure the values of thrres vary markedly with the wavelength combination. In Table 2 a bar indicates that for a given wavelength combination thrres = Resol (no threshold limitation is applied). For the combinations constituted of single wavelengths we found it useful to introduce the worst thrres value calculated for the combinations involving them (for example, for ApD, wavelength 1, thrres = 2.5 because the algorithm calculated thrres = 2.5 for the combination 1-4).

(iii) The value of *nf* does not substantially change with the wavelength combination, provided a correct solution is attained. This is because of the power of the automatic direct-space refinement routines: even an initial small substructure fragment is frequently expanded into a nearly complete substructure.

Table 3
For each test structure we give the cpu time necessary to find the first solution (min) using a XP1000 Compaq workstation when the wavelength combination chosen by the procedure is used.

Protein code	ApD	JIA	KPR	IDI	MDD	PSCP	Cyanase	Tm0665	TGEV	AEP	KSM
Cpu (min)	0.3	4.3	1	1.9	0.6	5	1.2	21	6.2	21.7	264.6

(iv) If a correct solution is attained the value of *nf* is very close to *na* for all the test structures.

Table 3 provides the cpu time (in minutes) necessary to find (for the selected combination) the first solution using a XP1000 Compaq workstation.

6. Conclusions

We have described two simple algorithms: one for limiting the data resolution and the second for predicting the most informative wavelength combinations in MAD techniques. A large set of crystal structures has been used to check their practical usefulness. The applications suggest the following final considerations.

(i) Diffraction data were collected at four wavelengths for ApD, JIA, cyanase and TGEV. For these the wavelength order in Table 2 is the following: 1, high-wavelength remote (H), 2, inflection (I), 3, peak (P), 4, low-wavelength remote (L). For all the test structures the high-wavelength data are less informative both when used alone (combinations 1 in the Table 2) and when combined with the other data (these combinations are confined to the right-hand side of Table 2).

(ii) For KPR, MDD, PSCP, Tmo665, AEP and KSM data were collected at three wavelengths (of type I, P, L in that order). The most informative combinations involve all three wavelengths.

(iii) Two-wavelength data were collected for IDI (of type I and P) and the best combination involves both.

(iv) In accordance with the observations above high-wavelength data are of little use for determining the substructure *via* MAD data (such a conclusion agrees well with the fact that f'' is quite small for high-wavelength data: therefore, $\Delta_{\rm ano}$ is more sensitive to experimental errors). Generally speaking, combinations of more wavelengths are more informative, in agreement with the indications arising from the joint probability distribution function theory.

It may be worthwhile mentioning that no experimental observations were preliminarily eliminated from the calculations on the basis of statistical criteria relying only the distribution of the experimental errors (*e.g.* when some reflections result in outliers). This choice may be justified as follows. Firstly, the distributions may be identical for informative and for non-informative data (see Fig. 1, where the frequency of $\Delta_{\rm ano}/\sigma_{\Delta{\rm ano}}$ for the four-wavelength data of ApD are shown). Secondly, the use of (3) to select the reflections to submit to the tangent formula (instead of choosing the reflections with the largest $\langle R_{oa}|\ldots\rangle$) is itself a very important filter, suggested by the application of the joint probability distribution method to MAD data. Indeed, anomalous or dispersive differences which are outliers for one or two wavelengths are frequently uncorrelated with the corresponding data collected at other wavelengths. In this case, large values of $\sigma_{R_{oa}}$ will indicate that data are uncorrelated and that the calculated $\langle R_{oa}|\ldots\rangle$ is not reliable.

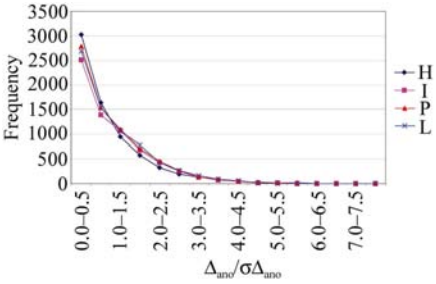

Figure 1
The frequency of $\Delta_{\rm ano}/\sigma_{\Delta{\rm ano}}$ for the four-wavelength data of ApD. L, low-wavelength remote; P, peak; I, inflection; H, high-wavelength remote.

We are indebted to S. K. Burley, Z. Dauter, K. Djinovic, R. Hilgenfeld, A. González, D. Matak, M. Walsh, C. Weeks and F. von Delft who kindly provided us with experimental data.

References

Anand, K., Palm, G. J., Mesters, J. R., Siddell, S. G., Ziebuhr, J. & Hilgenfeld, R. (2002). *EMBO J.* **21**, 3213–3224.
Bonanno, J. B., Edo, C., Eswar, N., Pieper, U., Romanowski, M. J., Ilyin, V., Gerchman, S. E., Kycia, H., Studier, F. W., Sali, A. & Burley, S. K. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 12896–12901.
Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (2003). *J. Appl. Cryst.* **36**, 1103.
Burla, M. C., Carrozzini, B., Caliandro, R., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* A**59**, 560–568.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* D**59**, 662–669.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2002). *Acta Cryst.* D**58**, 928–935.
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.
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.
Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). *Acta Cryst.* D**55**, 1568–1577.
Hendrickson, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11–21.
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**, 14493–14500.
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.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 11–16.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Von Delft, F., Inoue, T., Saldanha, S. A.,Ottenhof, H. H., Schmitzberger, F., Birch, L. M., Dhanara, V., Witty, M., Smith, A. G., Blundell, T. L. & Abell, C. (2003). *Structure*, **11**, 985–996.
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.