Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Enhancing MAD F_A data for substructure determination

Hongliang Xu

Mathematics Department, SUNY College at Buffalo, 1300 Elmwood Avenue, Buffalo, NY 14222, USA, and Hauptman–Woodward Medical Research Institute, 700 Ellicott Street, Buffalo, NY 14203, USA

Correspondence e-mail: xu@hwi.buffalo.edu

Received 1 April 2010
Accepted 30 June 2010

Heavy-atom substructure determination is a critical step in phasing an unknown macromolecular structure. Dual-space (*Shake-and-Bake*) recycling is a very effective procedure for locating the substructure (heavy) atoms using F_A data estimated from multiple-wavelength anomalous diffraction. However, the estimated F_A are susceptible to the accumulation of errors in the individual intensity measurements at several wavelengths and from inaccurate estimation of the anomalous atomic scattering corrections f' and f''. In this paper, a new statistical and computational procedure which merges multiple F_A estimates into an averaged data set is used to further improve the quality of the estimated anomalous amplitudes. The results of 18 Se-atom substructure determinations provide convincing evidence in favor of using such a procedure to locate anomalous scatterers.

1. Introduction

MAD (multiple-wavelength anomalous diffraction) is one of the primary techniques used to solve macromolecular structures. In a MAD experiment, two or more wavelengths are used to measure anomalous dispersion. The determination of a new protein structure is typically a two-step process. The first step is to locate the anomalous scatterers; the positions of the substructure atoms are then used as a bootstrap to initiate the phasing of the complete structure. Substructure atoms can be located using computational procedures that are based on either Patterson or direct methods. In either case, substructure intensities are estimated from MAD data by using either the so-called F_A formula, anomalous differences or dispersive differences, or various combinations thereof.

1.1. The analysis of MAD data

Let F_P, F_D and F_A be the structure-factor amplitudes of the native protein structure, the derivative structure and the substructure, and φ_P and φ_A be the native and substructure phases, respectively. Karle (1980) and Hendrickson *et al.* (1985) showed by algebraic analysis that for a given wavelength λ,

$$ {}^{\lambda}F_D^{\pm 2} = F_P^2 + a_\lambda F_A^2 + b_\lambda F_P F_A \cos(\Delta\varphi) \pm c_\lambda F_P F_A \sin(\Delta\varphi), \quad (1)$$

$$ a_\lambda = [({}^{\lambda}f')^2 + ({}^{\lambda}f'')^2]/f_0^2, \quad b_\lambda = 2({}^{\lambda}f')/f_0, \quad c_\lambda = 2({}^{\lambda}f'')/f_0,$$

$$ \Delta\varphi = \varphi_P - \varphi_{A,}$$

where f_0 is the normal atomic scattering factor, ${}^{\lambda}f'$ and ${}^{\lambda}f''$ are the anomalous scattering corrections (wavelength-dependent) and the '+' sign refers to reflection (h, k, l) and the '−' sign to reflection ($-h$, $-k$, $-l$).

doi:10.1107/S0907444910025783

Table 1
Selenium-substructure data sets used in this investigation.

	Selenium sites				
PDB code	Theoretical†	Actual‡	Space group	MAD resolution (Å)	Reference
1qcz	5	4	$I422$	2.0	Mathews *et al.* (1999)
1bx4	8	7	$P2_12_12$	2.3	Mathews *et al.* (1998)
1cb0	9	8	$P321$	2.2	Appleby *et al.* (1999)
1t5h	10	10	$P3_221$	2.5	Gulick *et al.* (2004)
1gso	13	13	$P2_12_12_1$	2.2	Wang *et al.* (1998)
1jxh	14	14	$P4_12_12$	3.1	Cheng *et al.* (2002)
1dbt	21	19	$P2_12_12$	2.5	Appleby *et al.* (2000)
1jen	24	22	$P2_1$	2.3	Ekstrom *et al.* (1999)
1jc4	28	24	$P2_1$	2.1	McCarthy *et al.* (2001)
1cli	28	28	$P2_12_12_1$	3.0	Li *et al.* (1999)
1a7a	32	30	$C222$	2.8	Turner *et al.* (1998)
1l8a	42	40	$P2_1$	2.6	Arjunan *et al.* (2002)
1e3m	48	45	$P2_12_12_1$	3.0	Lamers *et al.* (2000)
1hi8	50	50	$P3_2$	3.0	Butcher *et al.* (2001)
1m32	66	66	$P2_1$	2.6	Chen *et al.* (2002)
1dq8	68	60	$P2_1$	2.6	Istvan *et al.* (2000)
1e2y	70	60	$P2_1$	3.2	Alphey *et al.* (2000)
1eq2	70	70	$P2_1$	3.0	Deacon *et al.* (2000)

† Potential sites based on the amino-acid sequence. ‡ Number of sites reported in the published protein structure.

Each of the measurements of $F_{\rm D}^{\pm}$ at a given wavelength gives us two equations (1) and the different wavelengths may be treated as a system of simultaneous equations. For two or more wavelengths, (1) represents an overdetermined system of equations that can be analyzed to obtain values of $F_{\rm A}$, $F_{\rm P}$ and $\Delta\varphi$ for each reflection. Once $F_{\rm A}$ estimates are available, direct methods can be applied to locate anomalous scatterers. Subsequently, substructure phases, $\varphi_{\rm A}$, can be calculated from the refined substructure. The initial experimental map is then calculated from estimated $F_{\rm P}$ and protein structure phases $\varphi_{\rm P} = \Delta\varphi + \varphi_{\rm A}$.

1.2. $F_{\rm A}$ data estimation

In a typical MAD experiment, three wavelengths (λ_1, edge; λ_2, peak; λ_3, remote) are used to measure Bijvoet differences. The common procedure in protein crystallography is to use ALL measured data to estimate the substructure structure factors $F_{\rm A}$. In theory, these $F_{\rm A}$ values should be closer to the true substructure factors than those estimated from single-wavelength or two-wavelength anomalous dispersion data would be.

From a mathematical point of view, at least two wavelengths [four equations (1)] are needed to estimate the three unknown quantities $F_{\rm A}$, $F_{\rm P}$ and $\Delta\varphi$. When three-wavelength data are available, we could have four different $F_{\rm A}$ estimates from four different inputs: λ_1 and λ_2, λ_1 and λ_3, λ_2 and λ_3, and λ_1, λ_2 and λ_3 (common protocol). In theory, all the $F_{\rm A}$ estimates should be the same if the anomalous dispersion data are error-free. Unfortunately, the $F_{\rm A}$ values can be estimated only approximately and their accuracy depends on many factors, including the precision of intensity measurements at various wavelengths, the stability of the wavelengths during the whole data-collection session, the effect of crystal deterioration *etc.* Estimated $F_{\rm A}$ values can be divided into three classes: reliably estimated, overestimated or underestimated. If the number of significantly overestimated $F_{\rm A}$ is large, direct methods such as *Shake-and-Bake* will fail. From a statistical point of view, data averaged over four $F_{\rm A}$ estimates will reduce the number of overestimated reflections and should be better than any individual $F_{\rm A}$ estimates. To test our hypothesis, we carried out a series of computational experiments described in the next section.

2. Materials and methods

The relative merits of using different $F_{\rm A}$ estimates have been determined by a postmortem analysis of the *Shake-and-Bake* procedure (Miller *et al.*, 1994; Weeks & Miller, 1999) for 18 known SeMet protein substructures ranging in size from five to 70 Se sites in the asymmetric unit. Basic information such as the Protein Data Bank (PDB) code, number of Se atoms in the asymmetric unit (N_μ), space group and MAD data resolution is listed in Table 1 for these substructures. In each case, three wavelengths of anomalous dispersion data were available and the *SHELXC* program (Sheldrick, 2008) was used to calculate four sets of substructure $F_{\rm A}$ values using the four different kinds of input mentioned in the previous section. Each set of $F_{\rm A}$ data was then normalized using a modified version of the *SHELXD* program (Schneider & Sheldrick, 2002) to output the normalized substructure structure factors $E_{\rm A}$.

All sets of normalized substructure structure factors were truncated to 3 Å resolution, the remaining reflections were sorted in decreasing order according to their $E_{\rm A}$ values and the top $30N_\mu$ reflections were then selected to generate the $300N_\mu$ most reliable three-phase structure invariants for *SnB* (Weeks & Miller, 1999) applications. Samples of 1000 randomly positioned N_μ-atom trial structures were generated for each set of test data and subjected to $2N_\mu$ cycles of *SnB* dual-space refinement. Following refinement, the mean phase error (MPE) relative to the known substructure was determined for each trial structure and trials with MPE values of less than 30° were counted as solutions. In all cases, low MPE values were perfectly correlated with low values of the minimal function. The success rate, defined as the percentage of trial structures that converged to solution at the end of a fixed number of *SnB* cycles, provided an important indication of the quality of the chosen computational method. The latest version of the *SnB* program implementing the statistical minimal function (Xu & Hauptman, 2004; Xu *et al.*, 2005) was used to obtain the computational results reported in this paper.

3. Results

3.1. Results from individual $F_{\rm A}$ estimations

The success rates obtained from statistical *Shake-and-Bake* using each of the four $E_{\rm A}$ data sets for the 18 Se-atom test substructures are listed in Table 2 under the headings $E_{\rm A}(\lambda_1, \lambda_3)$, $E_{\rm A}(\lambda_2, \lambda_3)$, $E_{\rm A}(\lambda_1, \lambda_2)$ and $E_{\rm A}(\lambda_1, \lambda_2, \lambda_3)$, respectively, for the four different ways of generating $E_{\rm A}$ values.

Table 2
Comparative *SnB* success rates for the 18 Se-atom substructures with four different E_A data sets and their averaged data set $\langle E_A \rangle$.

Statistically higher success rates when compared with those of the $E_A(\lambda_1, \lambda_2, \lambda_3)$ data sets are highlighted in bold.

		Success rate (%)				
PDB code	Se sites	$E_A(\lambda_1, \lambda_3)$	$E_A(\lambda_2, \lambda_3)$	$E_A(\lambda_1, \lambda_2)$	$E_A(\lambda_1, \lambda_2, \lambda_3)$	$\langle E_A \rangle$
1qcz	4	14.2	13.7	15.6	14.7	16.8
1bx4	7	11.7	12.4	9.3	12.4	11.7
1cb0	8	**6.9**	1.8	**5.7**	2.9	3.5
1t5h	10	2.8	4.0	1.3	3.7	3.0
1gso	13	7.9	5.1	5.1	6.6	6.2
1jxh	14	0.0	0.0	0.0	0.0	**0.8**
1dbt	19	5.4	2.5	**9.0**	5.8	**8.1**
1jen	22	14.3	12.6	11.9	11.9	13.2
1jc4	24	24.9	23.8	12.7	32.7	32.2
1cli	28	**2.1**	1.2	0.0	0.7	**4.9**
1a7a	30	4.1	4.4	2.6	5.2	5.6
1l8a	40	5.7	2.0	0.0	12.9	**25.6**
1e3m	45	4.6	4.3	4.3	5.7	7.6
1hi8	50	0.0	1.3	**41.3**	2.9	**19.6**
1m32	66	1.3	1.4	0.0	2.6	**6.6**
1dq8	60	**16.9**	2.2	**19.3**	12.1	**16.3**
1e2y	70	0.5	0.5	2.3	3.9	2.8
1eq2	70	0.1	0.0	0.0	0.1	**0.8**
Average		6.9	5.2	7.4	7.6	10.3

Table 3
Comparative *SnB* success rates for the 18 Se-atom substructures with different E_A data sets and different reflection sortings..

Statistically higher success rates yielded from E_A and $\langle E_A \rangle$ reflection sortings are highlighted in bold.

		Success rate (%)			
PDB code	Reflection sorting	$E_A(\lambda_1, \lambda_3)$	$E_A(\lambda_2, \lambda_3)$	$E_A(\lambda_1, \lambda_2)$	$E_A(\lambda_1, \lambda_2, \lambda_3)$
1qcz	E_A	14.2	13.7	15.6	14.7
	$\langle E_A \rangle$	16.1	16.9	18.5	16.0
1bx4	E_A	11.7	12.4	9.3	12.4
	$\langle E_A \rangle$	12.9	11.9	11.4	12.2
1cb0	E_A	**6.9**	1.8	**5.7**	2.9
	$\langle E_A \rangle$	4.4	2.5	3.4	4.0
1t5h	E_A	2.8	4.0	1.3	3.7
	$\langle E_A \rangle$	3.4	3.1	**3.7**	2.5
1gso	E_A	7.9	5.1	5.1	6.6
	$\langle E_A \rangle$	6.7	5.9	6.5	6.6
1jxh	E_A	0.0	0.0	0.0	0.0
	$\langle E_A \rangle$	**0.6**	**1.1**	**1.3**	**0.9**
1dbt	E_A	5.4	2.5	9.0	5.8
	$\langle E_A \rangle$	**8.8**	**7.2**	9.0	**8.5**
1jen	E_A	14.3	12.6	11.9	11.9
	$\langle E_A \rangle$	14.0	13.4	13.7	14.7
1jc4	E_A	24.9	23.8	12.7	32.7
	$\langle E_A \rangle$	**32.7**	**32.0**	**31.9**	32.5
1cli	E_A	2.1	1.2	0.0	0.7
	$\langle E_A \rangle$	**5.0**	**5.0**	**4.7**	**5.3**
1a7a	E_A	4.1	4.4	2.6	5.2
	$\langle E_A \rangle$	5.7	5.3	**4.8**	5.2
1l8a	E_A	5.7	2.0	0.0	12.9
	$\langle E_A \rangle$	**25.0**	**28.6**	**29.8**	**30.0**
1e3m	E_A	4.6	4.3	4.3	5.7
	$\langle E_A \rangle$	6.2	**7.0**	**6.6**	**8.0**
1hi8	E_A	0.0	1.3	**41.3**	2.9
	$\langle E_A \rangle$	**10.8**	**17.6**	23.4	**18.1**
1m32	E_A	1.3	1.4	0.0	2.6
	$\langle E_A \rangle$	**6.5**	**5.2**	**4.7**	**6.3**
1dq8	E_A	16.9	2.2	19.3	12.1
	$\langle E_A \rangle$	18.9	**18.6**	17.1	**17.1**
1e2y	E_A	0.5	0.5	2.3	**3.9**
	$\langle E_A \rangle$	**2.1**	**1.5**	1.9	2.3
1eq2	E_A	0.1	0.0	0.0	0.1
	$\langle E_A \rangle$	**1.0**	**1.1**	**1.2**	**1.1**

We have observed from the first four columns of Table 2 that (i) five $E_A(\lambda_1, \lambda_3)$ data sets, two $E_A(\lambda_2, \lambda_3)$ data sets, four $E_A(\lambda_1, \lambda_2)$ data sets and eight $E_A(\lambda_1, \lambda_2, \lambda_3)$ data sets yielded the highest success rates, (ii) three $E_A(\lambda_1, \lambda_3)$ data sets, nine $E_A(\lambda_2, \lambda_3)$ data sets, 12 $E_A(\lambda_1, \lambda_2)$ data sets and two $E_A(\lambda_1, \lambda_2, \lambda_3)$ data sets yielded the lowest success rates and (iii) none of the four E_A estimates produced solutions for all 18 substructures. In fact, two $E_A(\lambda_1, \lambda_3)$ data sets (1jxh and 1hi8), two $E_A(\lambda_2, \lambda_3)$ data sets (1jxh and 1eq2), five $E_A(\lambda_1, \lambda_2)$ data sets (1jxh, 1cli, 1l8a, 1m32 and 1eq2) and one $E_A(\lambda_1, \lambda_2, \lambda_3)$ data set (1jxh) failed to yield solutions. Overall, $E_A(\lambda_1, \lambda_2, \lambda_3)$ data sets produced most of the highest success rates and the smallest number of failures. The results confirm why $E_A(\lambda_1, \lambda_2, \lambda_3)$ is chosen as a common protocol in protein crystallography to estimate the substructure structure factors.

3.2. Effects of the averaged F_A estimations

Since it is impossible to eliminate experimental errors completely or to predict which E_A estimates will fail to produce solutions, we can utilize statistical procedures to identify reflections with the most reliably estimated substructure structure factors. From a statistical point of view, each of the four data sets can be regarded as independent estimates of the substructure structure factors and the values of the averaged $\langle E_A \rangle$ data set should be more reliable than any of the individual estimates. To verify this hypothesis, all reflections in the averaged data set were sorted in decreasing order of the $\langle E_A \rangle$ values and the top $30N_\mu$ reflections were selected as input to statistical *Shake-and-Bake*. The success rates for the 18 Se-atom test substructures are listed in Table 2 under the heading $\langle E_A \rangle$. When comparing two success rates (x and y) obtained from two different data sets, y is statistically higher than x if $y \geq x + 2\sigma(|y - x|)$, where $\sigma(x)$ is the standard deviation calculated by Bernoulli's distribution, $\sigma(x) = [nx(1 - x)]^{1/2}$, where n is the number of trials, x is the success rate expressed as a fraction and $\sigma(|y - x|) = [\sigma^2(x) + \sigma^2(y)]^{1/2}$, and y is statistically lower than x if $y \leq x - 2\sigma(|y - x|)$; otherwise y is statistically equivalent to x. When compared with $E_A(\lambda_1, \lambda_2, \lambda_3)$ data sets for the 18 test substructures, $\langle E_A \rangle$ data sets yielded statistically higher success rates for eight test substructures (highlighted in bold in Table 2) and statistically equivalent success rates for the other ten test substructures. More importantly, $\langle E_A \rangle$ data sets yielded solutions for all 18 test substructures.

3.3. Effects of reflection selections

As shown in Table 2, zero success rates were observed with at least one of the four estimated data sets for six test substructures (1jxh, 1cli, 1l8a, 1hi8, 1m32 and 1eq2), but not with their averaged data sets. There are two major differences between the individual E_A data sets and the $\langle E_A \rangle$ data sets: (i) the amplitudes of the normalized substructure structure

Table 4
Investigation of E_A estimates using different two-wavelength combinations for the selected substructures.

Possible causes for zero or near-zero *SnB* success rates are highlighted in bold.

					Correlation coefficient with $\langle E_A \rangle$	
PDB code	E_A resolution† (Å)	No. of reflections	Reflection sorting	Percentage overestimates (%)	Overall	Overestimated
1jxh	**3.6 (3.1)**	420	$E_A(\lambda_1, \lambda_3)$‡	25	0.515	0.030
			$E_A(\lambda_2, \lambda_3)$‡	27	0.535	−0.158
			$E_A(\lambda_1, \lambda_2)$‡	28	0.528	−0.208
1cli	3.0 (3.0)	840	$E_A(\lambda_1, \lambda_3)$	41	0.387	0.100
			$E_A(\lambda_2, \lambda_3)$	37	0.461	0.022
			$E_A(\lambda_1, \lambda_2)$‡	39	**0.272**	−0.015
1l8a	2.8 (2.6)	1260	$E_A(\lambda_1, \lambda_3)$	47	0.242	0.002
			$E_A(\lambda_2, \lambda_3)$	41	0.210	0.065
			$E_A(\lambda_1, \lambda_2)$‡	47	**0.057**	−0.074
1hi8	3.0 (3.0)	1500	$E_A(\lambda_1, \lambda_3)$‡	**50**	**0.234**	0.050
			$E_A(\lambda_2, \lambda_3)$	36	0.360	0.022
			$E_A(\lambda_1, \lambda_2)$	38	0.331	−0.045
1m32	3.0 (2.6)	1980	$E_A(\lambda_1, \lambda_3)$	35	0.285	0.000
			$E_A(\lambda_2, \lambda_3)$	35	0.307	0.106
			$E_A(\lambda_1, \lambda_2)$‡	39	**0.148**	0.035
1eq2	**3.4 (3.0)**	2100	$E_A(\lambda_1, \lambda_3)$‡	44	0.260	0.012
			$E_A(\lambda_2, \lambda_3)$‡	39	0.313	0.048
			$E_A(\lambda_1, \lambda_2)$‡	40	0.259	0.086

† Values in parentheses indicate the MAD data resolutions input to the *SHELXC* program. ‡ Data sets that yielded zero or near-zero *SnB* success rates.

factors and (ii) the rankings of the reflections and thus the selection of the top reflections for *SnB* phasing. To investigate the possible cause of the zero success rates, we replaced the amplitudes of the reflections in the $\langle E_A \rangle$ data set with one of the four individually estimated E_A values, respectively, but kept the ranking of the reflections (no re-ranking). We then selected the top $30N_\mu$ reflections as input to statistical *Shake-and-Bake*. The corresponding success rates, listed in Table 3, showed significant improvement in success rates. When compared with E_A ranking for the 18 test substructures, $\langle E_A \rangle$ ranking yielded 35 statistically higher success rates (highlighted in bold in Table 3), 33 statistically equivalent success rates and only four statistically lower success rates. Furthermore, $\langle E_A \rangle$ ranking yielded statistically higher success rates in all but one case for these six selected test substructures (1jxh, 1cli, 1l8a, 1hi8, 1m32 and 1eq2), thereby indicating that the selection of the top reflections for phasing is in fact the major factor in successful *SnB* application. The large number of overestimated E_A amplitudes is the main reason for the low or near-zero success rates. The reflection ranking based on the $\langle E_A \rangle$ amplitudes, rather than the E_A amplitudes, effectively eliminates the overestimated reflections and thus significantly improves the success of the *Shake-and-Bake* applications.

3.4. Effects of measurement error

It has been shown in the previous sections that the presence of overestimated *E*s in the list of the highest E_A amplitudes leads to failure in the direct-methods application when customary *E* estimates are employed for substructure solution. Overestimated *E*s stem from measurement errors in the MAD data. In this section, we try to inspect the overestimated reflections for systematic trends and investigate the reasons for the zero *SnB* success rates for the substructures that yielded at least one *SnB* failure. Information such as the PDB code, data resolution (including both MAD data and E_A substructure data), number of reflections for *SnB* phasing and reflection sorting is listed in the first four columns of Table 4 for the six selected substructures. Using $\langle E_A \rangle$ as a reference, reflections that appear in the top $30N_\mu$ (where N_μ is the number of Se sites) list of individual E_A estimates but do not appear in the top list of the averaged data set, $\langle E_A \rangle$, are considered as overestimated reflections. For each set of two-wavelength estimates, the percentage of overestimated reflections and the correlation coefficient (CC) between individual estimates E_A and $\langle E_A \rangle$ for all $30N_\mu$ reflections and for overestimated reflections only are also listed in Table 4.

First of all, the percentage of overestimated *E*s ranges from 25 to 50% and there is no correlation between the overestimated *E*s and their averaged $\langle E_A \rangle$ (indicated by near-zero CC values). For substructures 1jxh and 1eq2 outliers in the 3.6–3.1 and 3.4–3.0 Å resolution shells, respectively, are rejected by the *SHELXC* program, indicating poor quality of the integrated intensities in the resolution shell (perhaps owing to ice rings). The main reason for *SnB* failures for these two substructures is the combination of low E_A data resolution and overestimated *E*s. For substructures 1cli, 1l8a, 1hi8 and 1m32, a relatively low overall CC (bold numbers in column 6 of Table 4) from one of the three combinations indicates that one of the three-wavelength MAD data sets might have unacceptable quality (perhaps owing to radiation damage). For example, the overall CC from substructure 1hi8 indicates that the data quality of $E_A(\lambda_1, \lambda_3)$ (CC = 0.236) is poorer than those of $E_A(\lambda_2, \lambda_3)$ (CC = 0.360) or $E_A(\lambda_1, \lambda_2)$ (CC = 0.331). The *SnB* success rates [0.0, 1.3 and 41.3% for $E_A(\lambda_1, \lambda_3)$, $E_A(\lambda_2, \lambda_3)$ and $E_A(\lambda_1, \lambda_2)$, respectively] clearly indicate that the remote-wavelength anomalous difference had an unacceptable quality, possibly owing to radiation damage since four-wavelength MAD data were measured and the remote-wavelength data were measured last.

4. Conclusions

A new statistics-based procedure for substructure phasing using F_A formulae has been proposed and tested on 18 Se-atom substructure examples. The procedure successfully identified overestimated reflections from F_A formulae to be a common cause of phasing failures and suggested effective ways to enhance the F_A estimation. The test results demonstrate that the improvements are significant, especially for those substructures previously deemed difficult to determine. Although the results were based on the *Shake-and-Bake* application, this new procedure can be applied to any direct-methods-based substructure determination.

This research was supported by NIH grant GM072023. We thank everyone who made his or her SeMet MAD data available to us as test data.

References

Alphey, M. S., Bond, C. S., Tetaud, E., Fairlamb, A. H. & Hunter, W. N. (2000). *J. Mol. Biol.* **300**, 903–916.

Appleby, T. C., Erion, M. D. & Ealick, S. E. (1999). *Structure*, **7**, 629–641.

Appleby, T. C., Kinsland, C. L., Begley, T. P. & Ealick, S. E. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 2005–2010.

Arjunan, P., Nemeria, N., Brunskill, A., Chandrasekhar, K., Sax, M., Yan, Y., Jordan, F., Guest, J. R. & Furey, W. (2002). *Biochemistry*, **41**, 5213–5221.

Butcher, S. J., Grimes, J. M., Makeyev, E. V. & Bamford, D. H. (2001). *Nature (London)*, **410**, 235–240.

Chen, C. C. H., Zhang, H., Kim, A. D., Howard, A., Sheldrick, G. M., Dunaway-Mariano, D. & Herzberg, O. (2002). *Biochemistry*, **41**, 13162–13169.

Cheng, G., Bennett, E. M., Begley, T. P. & Ealick, S. E. (2002). *Structure*, **10**, 225–235.

Deacon, A. M., Ni, Y. S., Coleman, W. G. Jr & Ealick, S. E. (2000). *Structure*, **8**, 453–462.

Ekstrom, J. L., Mathews, I. I., Stanley, B. A., Pegg, A. E. & Ealick, S. E. (1999). *Structure*, **7**, 583–595.

Gulick, A. M., Lu, X. & Dunaway-Mariano, D. (2004). *Biochemistry*, **43**, 8670–8679.

Hendrickson, W. A., Smith, J. L. & Sheriff, S. (1985). *Methods Enzymol.* **115**, 41–55.

Istvan, E. S., Palnitkar, M., Buchanan, S. K. & Deisenhofer, J. (2000). *EMBO J.* **19**, 819–830.

Karle, J. (1980). *Int. J. Quantum Chem. Symp.* **7**, 357–367.

Lamers, M. H., Perrakis, A., Enzlin, J. H., Winterwerp, H. H., De Wind, N. & Sixma, T. K. (2000). *Nature (London)*, **407**, 711–717.

Li, C., Kappock, T. J., Stubbe, J., Weaver, T. M. & Ealick, S. E. (1999). *Structure*, **7**, 1155–1166.

Mathews, I. I., Erion, M. D. & Ealick, S. E. (1998). *Biochemistry*, **37**, 15607–15620.

Mathews, I. I., Kappock, T. J., Stubbe, J. & Ealick, S. E. (1999). *Structure*, **7**, 1395–1406.

McCarthy, A. A., Baker, H. M., Shewry, S. C., Patchett, M. L. & Baker, E. N. (2001). *Structure*, **9**, 637–646.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Turner, M. A., Yuan, C. S., Borchardt, R. T., Hershfield, M. S., Smith, G. D. & Howell, P. L. (1998). *Nature Struct. Biol.* **5**, 369–376.

Wang, W., Kappock, T. J., Stubbe, J. & Ealick, S. E. (1998). *Biochemistry*, **37**, 15647–15662.

Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.

Xu, H. & Hauptman, H. A. (2004). *Acta Cryst.* A**60**, 153–157.

Xu, H., Weeks, C. M. & Hauptman, H. A. (2005). *Acta Cryst.* D**61**, 976–981.