Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Rapid and automated substructure solution by *Shake-and-Bake*

Hongliang Xu* and Charles M. Weeks

Hauptman–Woodward Medical Research Institute and Department of Structural Biology, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA

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

Received 29 August 2007
Accepted 13 November 2007

Direct methods of phase determination have played an important role in determining heavy-atom substructures from difference amplitudes of native–derivative crystal pairs or crystals containing anomalously scattering atoms. The minimal principle-based *Shake-and-Bake* procedure is one of the most successful direct methods for heavy-atom substructure determination. The computer program *SnB*, which implements the *Shake-and-Bake* procedure and is part of the protein structure-determination package *BnP*, has recently been optimized for rapid and automated substructure determination. Specifically, *SnB* has been upgraded with (i) a newly developed statistical minimal function for higher success rates, (ii) an optimal FFT grid size for dramatic cost-effectiveness improvement, (iii) a dynamic figure of merit for automatic substructure-solution detection and (iv) a strategy of alternation of anomalous differences with isomorphous dispersive differences for virtually guaranteed substructure solution.

1. Introduction

Macromolecular crystal structure determination has typically been a two-step process. When diffraction data from multiple chemically isomorphous or anomalously scattering crystals (Green *et al.*, 1954; Harker, 1956; Steitz, 1968) are available, one first locates the positions of heavy atoms from difference amplitudes arising from native–derivative crystal pairs or anomalous scattering atoms and then completes the phasing of the whole protein structure by using the heavy-atom substructure as a bootstrap. Although both Patterson and direct methods can effectively determine small substructures, direct methods tend to be more efficient for large substructure determinations.

Shake-and-Bake (DeTitta *et al.*, 1994; Weeks *et al.*, 1994; Weeks & Miller, 1999) is a direct-methods procedure that automatically and repetitively alternates reciprocal-space phase refinement either by using the tangent formula (Karle & Hauptman, 1956) or by reducing the value of the minimal function (Debaerdemaeker & Woolfson, 1983) with complementary real-space density modification to impose physical constraints. *Shake-and-Bake* belongs to the class of phasing methods known as 'multi-solution' or 'multi-trial' procedures (Germain & Woolfson, 1968). Multiple trial structures are created by using a random-number generator to assign initial coordinates and each trial is then subjected to the dual-space refinement process. Potential solutions are identified on the basis of minimal function values at the end of *SnB* refinement. The complete algorithm has been described in detail in several reviews (*e.g.* Weeks *et al.*, 2001; Sheldrick *et al.*, 2001).

doi:10.1107/S0907444907058659

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

	Selenium sites				
PDB code	Theoretical†	Actual‡	Space group	Resolution (Å)	Reference
1qcz	5	4	$I422$	1.50	Mathews *et al.* (1999)
1bx4	8	7	$P2_12_12$	2.25	Mathews *et al.* (1998)
1cb0	9	8	$P321$	2.20	Appleby *et al.* (1999)
1t5h	10	10	$P3_221$	2.50	Gulick *et al.* (2004)
1gso	13	13	$P2_12_12_1$	2.22	Wang *et al.* (1998)
1jxh	14	14	$P4_12_12$	2.30	Cheng *et al.* (2002)
1dbt	21	19	$P2_12_12$	2.49	Appleby *et al.* (2000)
1jen	24	22	$P2_1$	2.25	Ekstrom *et al.* (1999)
1jc4	28	24	$P2_1$	2.00	McCarthy *et al.* (2001)
1cli	28	28	$P2_12_12_1$	3.00	Li *et al.* (1999)
1a7a	32	30	$C222$	2.80	Turner *et al.* (1998)
1l8a	42	40	$P2_1$	2.60	Arjunan *et al.* (2002)
1e3m	48	45	$P2_12_12_1$	3.00	Lamers *et al.* (2000)
1hi8	50	50	$P32$	2.80	Butcher *et al.* (2001)
1gkp	54	54	$C222_1$	2.50	Abendroth *et al.* (2002)
1m32	66	66	$P2_1$	2.55	Chen *et al.* (2002)
1dq8	68	60	$P2_1$	2.33	Istvan *et al.* (2000)
1e2y	70	60	$P2_1$	3.20	Alphey *et al.* (2000)
1eq2	70	70	$P2_1$	2.91	Deacon *et al.* (2000)

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

Shake-and-Bake is a powerful procedure that is capable of providing *ab initio* solutions for structures containing as many as ~2000 independent non-H atoms (Frazão *et al.*, 1999) provided that accurate diffraction data have been measured to a resolution of 1.2 Å or better and several moderately heavy atoms (*e.g.* sulfur or iron) are present. It has also provided solutions for heavy-atom protein substructures containing as many as 160 Se atoms (von Delft *et al.*, 2003) provided that anomalous difference data have been measured to ~3.0 Å. The *Shake-and-Bake* algorithm has been implemented in the computer programs *SnB* (Miller *et al.*, 1994; Weeks & Miller, 1999) and *BnP* (Weeks *et al.*, 2002). It has also been implemented independently in the program *SHELXD* (Sheldrick, 1998; Schneider & Sheldrick, 2002).

In order to meet the high-throughput requirements of structural genomics projects, every aspect of the protein-phasing process, including substructure determination, has to be optimized. The main goal of this paper is to minimize the expected time to solution. This goal can be realised either by increasing the percentage of successful trial structures or by reducing the amount of computing time required for each trial. In the following sections, we will detail methods and strategies to increase the success rate by introducing a new type of minimal function (§3), by switching from the use of anomalous differences to isomorphous dispersive differences (§4), by decreasing running time *via* an optimized FFT grid (§5) and by automatic detection of the occurrence of the first solution (§6).

2. Materials and methods

The relative merits of different computational procedures have been determined by a postmortem analysis of different *Shake-and-Bake* variants using test data for 19 known 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, the number of Se atoms in the asymmetric unit (N), the space group and the data resolution for these substructures is listed in Table 1. In each case three wavelengths of anomalous dispersion data were available and the *DREAR* program (Blessing & Smith, 1999) was used to calculate the normalized difference structure-factor magnitudes $|E_\Delta|$ for both peak-wavelength anomalous difference data ($\mathrm{PK_{ano}}$) and dispersive data ($\mathrm{IP_{iso}}$) related to the differences between the inflection-point and high-energy remote wavelengths. For comparison, normalized structure-factor magnitudes $|E_A|$ were also computed for the substructures. Estimates of the MAD $|F_A|$ values (Karle, 1989; Hendrickson, 1991) were generated using the *SHELXC* program and normalized using a version of *SHELXD* modified to output $|E_A|$ values.

Each set of normalized magnitudes was truncated to 3 Å for substructure determination; the remaining reflections were sorted in decreasing order according to their $|E_\Delta|$ or $|E_A|$ values and the top $30N$ reflections were then selected to generate the $300N$ most reliable three-phase structure invariants. Samples of randomly positioned N-atom trial structures were generated for each set of test data and subjected to $2N$ cycles of *SnB* dual-space refinement using one of the methods described in §3. Following refinement, the mean phase error (MPE) relative to the known substructure was determined for each trial structure and trials with MPE values 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 converge to solution at the end of a fixed number of *Shake-and-Bake* cycles, provides an important indication of the quality of a particular computational method. However, this measurement does not take into account the computational effort (running time) needed to produce the solutions. The relative efficiency of different refinement methods can be compared on the basis of the cost-effectiveness CE,

$$\mathrm{CE} = \frac{3600S}{TCt}, \tag{1}$$

where T is the number of trial structures, C is the number of cycles per trial structure, S is the number of solutions produced by T such trials and t is the running time (in seconds) for one cycle of one trial. CE has units of solutions per hour and the expected time required to achieve the first solution is the reciprocal of CE.

3. Success-rate improvement: refinement method

If $E_H = |E_H|\exp(i\varphi_H)$ are the normalized structure factors of a crystal structure, then the so-called crystallographic phase problem is to determine the phases φ_H when the structure-factor magnitudes $|E_H|$ are available. In the minimal-principle method, the phase problem is formulated as a problem of constrained global minimization. The minimal function of the structure invariants, $\varphi_{HK} = \varphi_H + \varphi_K + \varphi_{-H-K}$, plays a critical

Table 2
Overall improvement of statistical over traditional *Shake-and-Bake* as measured by the statistical-to-traditional success-rate ratios for peak anomalous difference data (PK_{ano}) or dispersive difference data (IP_{iso}).

All of these ratios, except one, were greater than unity.

PDB code	Se sites	Data type	Ratio
1qcz	4	PK_{ano}	2.61
1bx4	7	PK_{ano}	1.67
1cb0	8	PK_{ano}	1.26
1t5h	10	PK_{ano}	1.28
1gso	13	IP_{iso}	1.12
1jxh	14	IP_{iso}	1.28
1dbt	19	PK_{ano}	1.84
1jen	22	PK_{ano}	1.15
1jc4	24	PK_{ano}	1.27
1cli	28	PK_{ano}	2.04
1a7a	30	PK_{ano}	1.85
1l8a	40	PK_{ano}	2.83
1e3m	45	PK_{ano}	2.34
1hi8	50	PK_{ano}	2.00
1gkp	54	IP_{iso}	2.00
1m32	66	PK_{ano}	1.73
1dq8	60	PK_{ano}	1.66
1e2y	70	IP_{iso}	0.78
1eq2	70	PK_{ano}	1.62

Table 3
Comparison of success rates (%) for 1000 *SnB* trials using PK_{ano}, IP_{iso} and F_A data for the 19 Se-atom substructures.

The highest success-rate values for each substructure are shown in bold.

PDB code	PK_{ano}	IP_{iso}	F_A
1qcz	13.5	11.8	**14.7**
1bx4	11.3	**19.9**	12.4
1cb0	4.7	**6.5**	3.9
1t5h	4.2	**6.3**	3.7
1gso	0.0	**13.9**	6.6
1jxh	1.0	**11.5**	0.0
1dbt	5.1	**8.2**	5.8
1jen	**12.0**	0.0	11.9
1jc4	28.8	0.0	**32.7**
1cli	3.9	**4.4**	0.7
1a7a	4.5	1.6	**5.1**
1l8a	3.0	**13.7**	12.9
1e3m	6.8	**7.0**	5.7
1hi8	26.6	0.0	**37.1**
1gkp	0.0	**2.2**	0.0
1m32	4.5	**28.1**	2.7
1dq8	**24.2**	10.8	12.1
1e2y	0.0	**10.2**	3.5
1eq2	**3.4**	0.0	0.1

role in the phase-refinement portion of the *Shake-and-Bake* cycle. Successful applications of *Shake-and-Bake* to structure determination depend on the formulation of minimal functions and their radius of convergence.

3.1. Cosine minimal function

Traditionally, *Shake-and-Bake* employs a probabilistic based cosine minimal function (DeTitta *et al.*, 1994),

$$R(\varphi) = \left(\sum_{H,K} A_{HK}\right)^{-1} \sum_{H,K} A_{HK}\left[\cos(\varphi_{HK}) - \frac{I_1(A_{HK})}{I_0(A_{HK})}\right]^2, \quad (2)$$

where $A_{HK} = 2N^{-1/2}|E_H E_K E_{H+K}|$, N is the number of non-H atoms (or substructure heavy-atom sites) in the asymmetric unit and I_1/I_0 is the ratio of modified Bessel functions. The cosine minimal function measures the least-squares difference between the cosine values of the structure invariants, calculated using a set of trial phases, and the theoretically estimated values of the same invariants. This minimal function serves as the foundation of traditional *Shake-and-Bake*. Several different types of probabilistic based minimal functions have been proposed (Hauptman *et al.*, 1999; Xu *et al.*, 2002).

3.2. Statistical minimal function

A new type of minimal function, termed the statistical minimal function (Xu & Hauptman, 2004), was formulated on the basis of empirical observation of the distribution of the three-phase structure invariants. Let $I = [-r, r]$ be an arbitrary interval, N_I be the number of triplet invariants whose values lie in I and N_{T} be the total (fixed) number of triplet invariants. The statistical minimal function is then simply defined as

$$m(\varphi) = 1 - (N_I/N_{\text{T}}). \quad (3)$$

It was conjectured and subsequently experimentally confirmed that the minimal function reaches its constrained global minimum when all phases are equal to their true values. The statistical minimal function serves as the foundation of a corresponding statistical *Shake-and-Bake*.

Applications of both traditional and statistical *Shake-and-Bake* were made to 1000 trial structures for each of the 19 test substructures using a modified version of the computer program *SnB* (Weeks & Miller, 1999). Success rates were computed for both methods and the statistical-to-traditional success-rate ratios are reported in the column labeled 'Ratio' in Table 2. Of the 19 test cases, 18 difference data sets yielded ratios that were greater than unity and 12 ratios were greater than 1.5. These results clearly illustrate the overall superiority of statistical *Shake-and-Bake*. As a consequence of such dramatic improvement, statistical *Shake-and-Bake*, along with its default statistical interval $I = [-r, r]$ with $r = \min\{9.14\ln(N) + 55.3°, 90°\}$ (Xu *et al.*, 2005), will be the default refinement method used throughout the remainder of this paper.

4. Success-rate improvement: difference data

When three wavelengths of anomalous dispersion data are available, one has at least three choices of different ways of determining the substructure [*i.e.* by using (i) peak-wavelength anomalous difference data (PK_{ano}), (ii) dispersive difference data (IP_{iso}) and (iii) MAD F_A data]. Success rates obtained from statistical *Shake-and-Bake* using each of these three data types for the 19 Se-atom test substructures are listed in Table 3. Firstly, three of the peak anomalous (PK_{ano}) data sets, 12 of the dispersive (IP_{iso}) data sets and four of the MAD F_A data sets yielded the highest success rates (bold numbers in Table 3). Secondly, none of the three data types produced solutions for all 19 substructures. In fact, three PK_{ano} data sets (1gso, 1gkp and 1e2y), four IP_{iso} data sets

Table 4
Effects of different Fourier grid sizes (1.0, 1.5 and 2.0 Å) on cost-effectiveness using statistical *Shake-and-Bake* for the 19 Se-atom substructures.

The highest cost-effectiveness values for an SGI R10000 workstation are shown in bold.

		Grid size (Å)		
PDB code	Data type	1.0	1.5	2.0
1qcz	PK_{ano}	90.3	144.5	**343.3**
1bx4	PK_{ano}	80.4	255.4	**394.1**
1cb0	PK_{ano}	23.7	46.0	**70.2**
1t5h	PK_{ano}	13.1	36.0	**58.2**
1gso	IP_{iso}	44.8	**161.0**	113.8
1jxh	IP_{iso}	15.1	75.7	**244.3**
1dbt	PK_{ano}	5.5	22.6	**22.7**
1jen	PK_{ano}	25.6	56.9	**101.3**
1jc4	PK_{ano}	61.1	130.7	**156.1**
1cli	PK_{ano}	1.5	3.2	**10.7**
1a7a	PK_{ano}	0.8	2.1	**9.5**
1l8a	PK_{ano}	0.9	**2.9**	2.4
1e3m	PK_{ano}	0.8	2.3	**6.2**
1hi8	PK_{ano}	3.7	8.3	**15.0**
1gkp	IP_{iso}	0.1	0.1	**0.3**
1m32	PK_{ano}	0.5	0.9	**2.2**
1dq8	PK_{ano}	3.6	**18.7**	17.1
1e2y	IP_{iso}	1.6	3.0	**10.2**
1eq2	PK_{ano}	0.2	0.6	**2.3**

(1jen, 1jc4, 1hi8 and 1eq2) and two F_A data sets (1jxh and 1gkp) failed to yield solutions. The possible causes of these failures were investigated by applying statistical *Shake-and-Bake* to error-free PK_{ano} data generated for 1gso, error-free IP_{iso} data generated for 1jc4 and error-free F_A data generated for 1jxh using the known atomic coordinates and the program *EGEN* (R. Blessing, personal communication). This study revealed that the success rates for these three error-free data sets were 11.8, 23.7 (Xu *et al.*, 2005) and 0.9%, respectively. Therefore, experimental error was in fact the cause of the zero success rates.

Since it is impossible to eliminate experimental errors completely or to predict which difference data sets will fail to produce solutions, it is important to devise a strategy that will ensure that whenever solutions exist they will be found and found as quickly as possible. Therefore, since errors in the anomalous and dispersive data sets are likely to be independent (in the three cases that PK_{ano} data sets failed to yield solutions, their companion IP_{iso} data sets produced success rates of 13.9, 2.2 and 10.2%, respectively; in the four cases that IP_{iso} data sets failed to yield solutions, their companion PK_{ano} data sets produced success rates of 12.0, 28.8, 26.6 and 3.4%, respectively), it is a good idea to process some trials using each type of data (PK_{ano} and IP_{iso}) in order to produce at least one solution in a limited number of trial structures and the latest version of the computer program *BnP* has adopted this strategy. In a single-processor environment, 500 *SnB* trials are refined using the PK_{ano} data and if these trials fail to produce a solution an additional 500 *SnB* trials are refined using the IP_{iso} data. In a multiprocessing environment, one or more independent multi-trial *SnB* refinement jobs can be executed simultaneously for both the PK_{ano} data and the IP_{iso} data. If a solution is detected automatically for one of these jobs (see §6), the other *SnB* refinement jobs can then be terminated.

5. Cost-effectiveness improvement: using an optimal FFT grid

One of the ways to maximize cost-effectiveness (CE) is to reduce *SnB* running time, in particular the CPU time required to complete one refinement cycle for one trial structure. *SnB* cycle time consists of the times for (i) phase refinement using the parameter-shift procedure to reduce the value of a minimal function, (ii) fast Fourier tranformation (FFT) to compute an electron-density map, (iii) density modification by peak picking or low-density elimination and (iv) an inverse FFT or structure-factor calculation to compute modified structure factors. The FFT time for calculating an electron-density map (the most time-consuming portion of the *SnB* cycle) depends heavily on the number of points at which the electron density is computed (*i.e.* on the grid size). In regular *SnB* applications to Se-atom substructure determinations (including those described in previous sections), the default grid size has traditionally been chosen to be one third of the data resolution (typically 1.0 Å). In this study, the effects of various grid sizes on success rate and running time (and therefore on cost-effectiveness) have been studied for the 19 Se-atom test substructures using statistical *Shake-and-Bake*. The results clearly demonstrate that using a coarse grid (2.0 Å) instead of a fine grid (1.0 Å) decreases success rates only modestly while greatly decreasing the running time and therefore significantly improves the cost-effectiveness, which (Table 4) ranges from a factor of 2.5 (1gso and 1jc4) to more than 10 (1jxh, 1a7a and 1eq2).

Using a coarse grid in *SnB* will decrease the quality of substructure solutions, resulting in higher minimal function (R_{min}) and crystallographic *R* (R_{cryst}) values, a narrowed gap between the R_{min} values for solutions and nonsolutions and a smaller number of accurately identified Se-atom sites. This will make it very difficult to identify potential solutions based on the automatic solution-detection test (§6). However, this difficulty can be overcome by adding one additional cycle of *SnB* refinement using a fine grid. The quality of the solutions, compared on the basis of the mean phase error (MPE) and listed in Table 5, is the same as those obtained when a fine grid is used for all cycles.

6. Cost-effectiveness improvement: automatic solution detection

Since *Shake-and-Bake* is implemented in the form of a multiple-trial procedure, it is essential to have some means of distinguishing the refined trial structures that are solutions from those that are not. It is important that solutions can be recognized quickly and automatically, that at least one solution can be identified in every case where some solutions are present and that nonsolutions never be falsely identified as solutions. The minimal function (R_{min}) and the crystallographic *R* factor (R_{cryst}) are two figures of merit (FOMs) that

Table 5
Comparison of mean phase errors (°) produced by (i) $2N - 1$ refinement cycles with a coarse grid (2 Å), (ii) protocol (i) plus an additional cycle with a fine grid (1 Å) and (iii) $2N$ refinement cycles with a fine grid (1 Å).

Protocol	(i)	(ii)	(iii)
1qcz	30.6	3.5	3.5
1bx4	24.7	17.4	12.0
1cb0	28.1	17.7	17.0
1t5h	20.7	7.7	8.4
1gso	32.1	8.0	7.2
1jxh	16.4	12.2	12.2
1dbt	25.9	15.2	15.2
1jen	38.7	19.9	19.3
1jc4	29.6	14.8	12.6
1cli	28.4	17.9	16.8
1a7a	21.8	5.9	5.3
1l8a	30.0	15.5	11.1
1e3m	23.5	15.2	12.5
1hi8	31.8	23.1	22.5
1gkp	21.6	14.8	14.8
1m32	39.0	21.1	26.6
1dq8	39.1	29.4	28.4
1e2y	26.1	20.5	19.1
1eq2	28.7	18.3	16.0

are computed at the end of *SnB* refinement for every trial structure and they are useful for making decisions about whether or not a solution has been found. Typically, the presence of a bimodal distribution of $R_{\rm min}$ values (based on a predetermined number of trial structures and presented in the form of a histogram) suggests that potential solutions have been found. The existence of solutions is confirmed by corresponding low values of $R_{\rm cryst}$ and the trial structure with the lowest $R_{\rm min}$ value is then selected as the solution. Since recognition of a biomodal histogram requires numerous trial structures and manual intervention, it is not suitable for automated applications.

Although the relative values of each FOM for any given substructure clearly distiguish solutions from nonsolutions, the exact range of FOM values for both solutions and nonsolutions varies from structure to structure. It is often the case that the value of an FOM for a nonsolution of a small substructure is less than that for a solution of a large substructure. This makes it impossible to specify a single cutoff value for each FOM that will discriminate solutions from nonsolutions for all structures or all substructures. However, for any given substructure it is observed that the values of each FOM for solutions are significantly lower than those for nonsolutions and no solution with a value of $R_{\rm cryst} \geq 0.33$ has ever been observed. Therefore, let R_j denote the FOM for the jth trial structure, $1 \leq j \leq T$, where T is the total number of *SnB* trial structures. Dynamic average values of $R_{\rm min}$ and its standard deviation can then be designed to detect automatically whether or not a solution is present among the completed trial structures. For example, if a solution has not been detected after m *SnB* trial structures (nonsolution sample $m \geq 5$), the average values of an FOM, $\overline{R}_m = (1/m)\sum_{j=1}^{m} R_j$, and its standard deviation, $\sigma_m = [1/(m-1)]\sum_{j=1}^{m}(R_j - \overline{R}_m)^2]^{1/2}$, are calculated and the next *SnB* trial structure is subject to the following solution test. If $R_{m+1} < \overline{R}_m - 3\sigma_m$, then the $(m+1)$th trial structure is a solution; otherwise, repeat the procedure replacing m by $m + 1$. Of course, in order to detect an early occurrence of a solution, the identity of the trial structure with the smallest $R_{\rm min}$ value among the *SnB* trial structures having $R_{\rm cryst}$ values less than 0.33 is tracked and automatically subjected to the solution test when the nonsolution ($R_{\rm cryst} \geq 0.33$) sample equals 5. When the solution test was applied to all test structures in Table 1, the first solution of every structure having at least one solution was identified correctly.

7. Conclusions

Four methods of improving *Shake-and-Bake* substructure determination have been proposed and tested on 19 Se-atom examples. These methods include (i) a new statistical minimal function that increases the percentage of trial structures that go to solution, (ii) a method for circumventing measurement errors in MAD data by using peak anomalous and dispersive difference data independently, (iii) minimizing computational time by using an optimum Fourier grid size for the density-modification step in the *SnB* cycle and (iv) an effective procedure for detecting solutions as soon as possible. These improvements have been implemented in the latest version of the computer program *BnP*, which can be downloaded from the web site http://www.hwi.buffalo.edu/BnP/. The use of the statistical minimal function and the optimum Fourier grid size result in more than a 40-fold reduction in the computing time required to solve the 160-site selenomethionine substructure of *Escherichia coli* ketopantoate hydroxymethyltransferase (KPHMT). The KPHMT substructure was first solved using an older version of *SnB* (von Delft *et al.*, 2003).

This research was supported by NIH grants EB002057 and GM072023. We thank everyone who made his or her SeMet MAD data available to us as test data.

References

Abendroth, J., Niefind, K. & Schomburg, D. (2002). *J. Mol. Biol.* **320**, 143–156.

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.

Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.

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.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.

Delft, F. von, Inoue, T., Saldanha, S. A., Ottenhof, H. H., Schmitzberger, F., Birch, L. M., Dhanaraj, V., Witty, M., Smith, A. G., Blundell, T. L. & Abell, C. (2003). *Structure*, **11**, 985–996.

DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.

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

Frazão, C., Sieker, L., Sheldrick, G. M., Lamzin, V., LeGall, J. & Carrondo, M. A. (1999). *J. Biol. Inorg. Chem.* **4**, 162–165.

Germain, G. & Woolfson, M. M. (1968). *Acta Cryst.* B**24**, 91–96.

Green, D. W., Inoram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 287–307.

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

Harker, D. (1956). *Acta Cryst.* **9**, 1–9.

Hauptman, H. A., Xu, H., Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* A**55**, 891–900.

Hendrickson, W. (1991). *Science*, **254**, 51–58.

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

Karle, J. (1989). *Acta Cryst.* A**45**, 303–307.

Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.

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.

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

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.

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. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer Academic Publishers.

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *International Tables for Crystallography*, Vol. F, edited by M. G. Rossmann & E. Arnold, pp. 333–345. Dordrecht: Kluwer Academic Publishers.

Steitz, T. A. (1968). *Acta Cryst.* B**24**, 504–507.

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**, 1564–15662.

Weeks, C. M., Blessing, R. H., Miller, R., Mungee, R., Potter, S. A., Rappleye, J., Smith, G. D., Xu, H. & Furey, W. (2002). *Z. Kristallogr.* **217**, 686–693.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.

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

Weeks, C. M., Sheldrick, G. M., Miller, R., Usón, I. & Hauptman, H. A. (2001). *Advances in Structure Analysis*, edited by R. Kužel & J. Hašek, pp. 37–64. Prague: Czech & Slovak Crystallographic Assocation.

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

Xu, H., Hauptman, H. A. & Weeks, C. M. (2002). *Acta Cryst.* D**58**, 90–96.

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