Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Optimizing *DREAR* and *SnB* parameters for determining Se-atom substructures

P. Lynne Howell,[a,b]* Robert H. Blessing,[c] G. David Smith[c,d] and Charles M. Weeks[c]

[a]Structural Biology and Biochemistry, Research Institute, Hospital for Sick Children, 555 University Avenue, Toronto M5G 1X8, Ontario, Canada, [b]Department of Biochemistry, Faculty of Medicine, University of Toronto, Medical Sciences Building, Toronto M5S 1A8, Ontario, Canada, [c]Hauptman–Woodward Medical Research Institute, 73 High Street, Buffalo, New York 14203, USA, and [d]Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, New York 4263, USA

Correspondence e-mail: howell@sickkids.on.ca

Received 2 November 1999
Accepted 23 February 2000

The determination of the anomalous scattering substructure is the first essential step in any successful macromolecular structure determination using the multiwavelength anomalous diffraction (MAD) technique. The diff*E* method of calculating difference *E*s in conjunction with *SnB* has had considerable success in determining large Se-atom substructures. An investigation of the parameters used in both the data-reduction and error-analysis routines (*DREAR*) as well as the *SnB* phasing process itself was undertaken to optimize these parameters for more efficient use of the procedure. Two sets of selenomethionyl *S*-adenosylhomocysteine hydrolase MAD data were used as test data. The elimination of all erroneously large differences prior to phasing was found to be critical and the best results were obtained from accurate highly redundant intensity measurements. The high-resolution data collected in the typical MAD experiment are sufficient, but the inclusion of low-resolution data below 20 Å improved the success rate considerably. Although the best results have been obtained from single-wavelength peak anomalous diffraction data alone, independent *SnB* analysis of data measured at other wavelengths can provide confirmation for questionable sites.

1. Introduction

In the past decade, there has been an exponential increase in the number of protein structures determined using multiwavelength anomalous diffraction (MAD) data. This increase is the result of the widespread use of low-temperature data-collection methods, the increased availability of suitable synchrotron beamlines and the ability to use standard molecular biological tools to systematically incorporate the anomalous scatterer selenium, in the form of selenomethionine, into proteins. Two different methods have been used to solve protein structures using MAD data. The first approach uses the algebraic formalism developed by Hendrickson (1985) based on an original formulation by Karle (1980). In this method, amplitudes for the normal scattering component of the anomalous substructure ($|F_A|$), amplitudes for the normal scattering due to all atoms ($|F_T|$) and the phase difference between them ($\Delta\varphi = \varphi_T - \varphi_A$) are calculated. The second approach treats the MAD data as a special case of multiple isomorphous replacement (MIR; Ramakrishnan & Biou, 1997). Regardless of the approach chosen, the first step in the structure determination is always the determination of the positions of the anomalous scatterers. In the past, these atoms were typically found from peaks in a Patterson map. These maps are calculated using the $|F_A|$s in the case of the algebraic formulation and either the anomalous

($||F_{+H}| - |F_{-H}||$) differences at the peak wavelength (λ_2) or the dispersive 'isomorphous' ($||F_{\lambda_1}| - |F_{\lambda_1}||$) differences between the inflection point ('derivative', λ_1) and the remote ('native', λ_3) data. While traditional Patterson techniques are sufficient to locate the Se-atom positions when the number of Se atoms is small, considerable difficulties are encountered when trying to locate larger numbers of Se atoms. Since the natural abundance of methionine in proteins is about 2%, as the size of the protein increases, so will the number of methionine residues and the number of Se-atom positions that need to be located. Recently, several alternative Patterson search techniques have been developed and incorporated into the programs *CNS* (Brunger *et al.*, 1998) and *SOLVE* (Terwilliger *et al.*, 1987). These algorithms have extended the number of Se-atom positions that can be found using Patterson techniques to at least 30 atoms and, in one case, as many as 52 (Janson, Smith & Terwilliger, personal communication).

An alternative approach to the location of the anomalous scatterers is to treat the problem as a small-molecule structure determination and employ direct-methods approaches to find the substructure (Mukherjee *et al.*, 1989). In 1995, *MULTAN* (Germain *et al.*, 1971) was used in conjunction with *E* magnitudes derived from the $|F_A|$ values to determine the position of four Se atoms in the biotinyl domain of acetyl-coenzyme A carboxylase (Athappilly & Hendrickson, 1995). The *MULTAN* program had also been used previously to verify the positions of three Se atoms obtained by Patterson methods in the structure determination of ribonuclease H (Yang *et al.*, 1990). However, these Se-atom substructures are small compared with the 15 Se-atom substructure used to solve the structure of bacteriophage T7 DNA replication complex (Doublie *et al.*, 1998). In this case, the dispersive differences ($||F_{\lambda_1}| - |F_{\lambda_3}||$) and the program *SHELXS*86 (Sheldrick, 1990) were used. While these applications of classical direct-method techniques showed that Se-atom substructures could be treated as small molecules, it was not obvious that these techniques could be routinely applied to larger substructures.

Second generation direct-methods programs such as *SnB* (Miller *et al.*, 1994; Weeks & Miller, 1999) have been successful in determining difficult small-molecule structures (Weeks *et al.*, 1993) and even small macromolecular structures for which atomic resolution data are available (Anderson *et al.*, 1996; Deacon *et al.*, 1998; Prive *et al.*, 1999; Smith *et al.*, 1997; Weeks *et al.*, 1995). Smith and colleagues (Blessing & Smith, 1999; Smith *et al.*, 1998) have developed a novel data-reduction procedure that combines the use of *SnB* with renormalized anomalous difference *E* magnitudes (diff*E*s) in order to determine Se-atom substructures. The first application of this method to an unknown protein was the *de novo* structure determination of *S*-adenosylhomocysteine (AdoHcy) hydrolase (Turner *et al.*, 1998), in which 30 out of 30 Se atoms were located from peak data (λ_2) alone. This method has subsequently been used to solve even larger Se-atom substructures. For example, 48 out of 56 Se atoms were found during the investigation of the EphB2 receptor SAM domain (Thanos *et al.*, 1999) and 65 out of 70 Se atoms were located in the study of ADP-L-glycero-D-mannoheptose 6-epimerase (Deacon *et al.*, 1999). An alternative procedure that utilized *E* magnitudes derived from $|F_A|$s has also been used in conjunction with *SnB* to solve a large Se-atom substructure (21 out of 24 Se atoms found; Smith & Krahn, personal communication).

Given the ease of use and the success of the diff*E* method (Blessing & Smith, 1999; Smith *et al.*, 1998) in conjunction with *SnB* (Weeks & Miller, 1999), an investigation of the parameters used in data reduction and phasing was undertaken in order to understand which were the most critical. In this study, a variety of different parameters have been systematically examined (Fig. 1) using two sets of MAD data collected from two different crystals of AdoHcy hydrolase. The results of this analysis have enabled default parameter values as well as a set of general guidelines to be proposed for this procedure.

2. Methods

2.1. Data collection

Two MAD data sets were collected on two different crystals of selenomethionyl-incorporated AdoHcy hydrolase at Station X12-C, National Synchrotron Light Source, Brookhaven National Laboratory. Both sets of data were collected at 100 K using inverse-beam geometry. Each data set was measured as a single wedge of data. The crystals belong to space group *C*222, with unit-cell parameters a = 91.93, b = 168.02, c = 137.77 Å. AdoHcy hydrolase is a homotetramer of 55 kDa subunits and crystallizes with a dimer in the

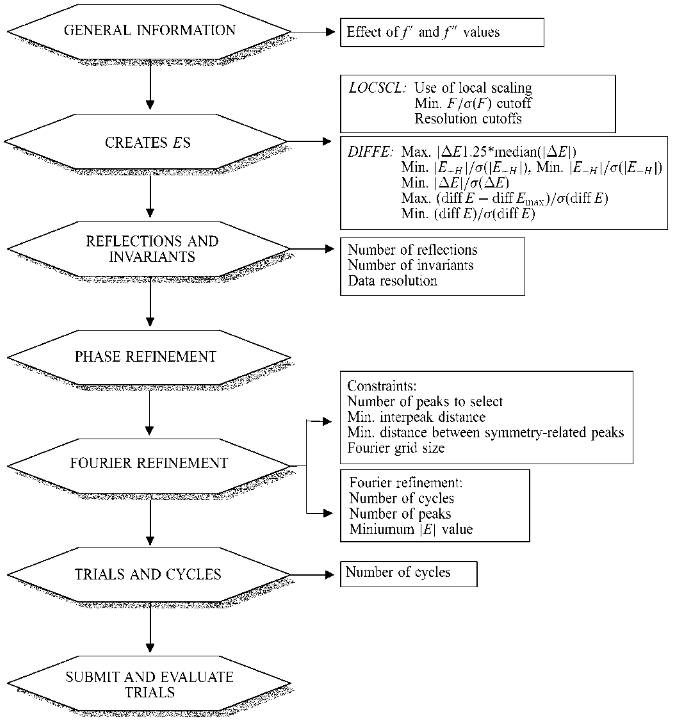

GENERAL INFORMATION
Effect of f' and f'' values
LOCSCL: Use of local scaling
Min. $F/\sigma(F)$ cutoff
Resolution cutoffs
CREATES ES
DIFFE: Max. $|\Delta E 1.25$*median$(|\Delta E|)$
Min. $|E_{-H}|/\sigma(|E_{-H}|)$, Min. $|E_{-H}|/\sigma(|E_{-H}|)$
Min. $|\Delta E|/\sigma(\Delta E)$
Max. (diff E − diff $E_{\max}$)/σ(diff E)
Min. (diff E)/σ(diff E)
REFLECTIONS AND INVARIANTS
Number of reflections
Number of invariants
Data resolution
PHASE REFINEMENT
Constraints:
Number of peaks to select
Min. interpeak distance
Min. distance between symmetry-related peaks
Fourier grid size
FOURIER REFINEMENT
Fourier refinement:
Number of cycles
Number of peaks
Miniumum $|E|$ value
TRIALS AND CYCLES
Number of cycles
SUBMIT AND EVALUATE TRIALS

Figure 1
Schematic diagram of the *SnB* menus and the parameters examined in this study. The parameters are aligned according to the menu on the graphical user interface (GUI) on which they can be found.

Table 1
Data-reduction statistics.

	SAH1			SAH2		
	Edge	Peak	Remote	Edge	Peak	Remote
Wavelength	0.9789	0.9784	0.95	0.9789	0.9784	0.95
f', f''	−9.52, 3.15	−7.35, 5.92	−2.74, 3.61	−9.52, 3.15	−7.35, 5.92	−2.74, 3.61
Resolution, min. (Å)	50	50	50	20	20	20
Resolution, max. (Å)	2.8	2.8	2.8	2.6	2.6	2.6
Total data†	345476	347188	350530	165143	165752	162628
Unique data†	50276	50469	50613	57712	57731	57985
Redundancy†	6.9	6.9	6.9	2.9	2.9	2.8
Completeness‡ (%)	99.3 (97.7)	99.4 (97.8)	99.5 (99.5)	91.3 (70.3)	91.3 (70.7)	91.4 (77.2)
$F^2 > 3\sigma(F^2)$‡ (%)	93.1 (82.3)	91.8 (79.9)	91.3 (80.2)	72.3 (32.9)	70.6 (30.0)	71.1 (35.5)
$R_{\rm merge}$§ (%)	6.8	7.1	7.4	5.2	5.3	5.3
$(\langle n\rangle/\langle n\rangle - 1)^{1/2}\,R_{\rm merge}$ (%)	7.8	7.7	8.0	6.4	6.5	6.5
Average $F^2/\sigma(F^2)$‡	23.6 (12.3)	22.2 (11.3)	21.6 (11.0)	17.1 (4.3)	16.2 (3.8)	15.7 (3.9)

† All data were processed with the anomalous pairs kept separate and scaled independently. The redundancy is therefore the average redundancy for each reflection of an anomalous pair. ‡ The figures quoted for overall completeness are for data between 50 and 2.8, and 20 and 2.6 Å for SAH1 and SAH2, respectively. The value in parentheses is the value for the final 0.1 Å resolution shell (*i.e.* 2.9–2.8 Å for SAH1 and 2.7–2.6 Å for SAH2). § $R_{\rm merge} = \sum\sum|I_i - \langle I\rangle|/\sum I_i$, where $\langle I\rangle$ is the average of equivalent reflections and the sum is extended over all measured observations for all unique reflections.

asymmetric unit. Owing to cleavage of the N-terminal methionine, as determined by N-terminal sequence analysis, each monomer contains 15 methionine residues, giving a total of 30 in the asymmetric unit. For the first set of data (SAH1), 369 images of 1° $\Delta\varphi$ oscillations were collected (*i.e.* 369° of data). These data were 99.5% complete (Table 1) and had an average measurement redundancy of 6.9 for the anomalous data (or an average redundancy of ∼14 if the Bijvoet pairs were merged). Since the diffraction from the first crystal did not extend beyond 2.8 Å, a second set of data was collected. For this data (SAH2), an oscillation range of 0.75° was used and 328 images were collected (*i.e.* 246° of data). This yielded an average redundancy of 2.8 for the anomalous data and 91% completeness. The length of the exposure was identical for both sets of data; the crystal-to-detector distances were 137 and 142 mm for SAH1 and SAH2, respectively. During the data processing of SAH2 using *DENZO*/*SCALEPACK*, it was discovered that the mosaicity of this crystal was 0.8°. All the reflections in the SAH2 data set were therefore measured as partial reflections.

All data were initially processed using the *DENZO*/*SCALEPACK* program (Otwinowski & Minor, 1997), treating the Bijvoet pairs as independent measurements. Table 1 presents the *DENZO*/*SCALEPACK* data-processing statistics. The merged *SCALEPACK* output data for each wavelength were processed further using six of the programs in the *DREAR* package (Blessing, 1989). First, *SORTAV* (Blessing, 1997*b*) was used to reformat the *SCALEPACK* merged data. Bayesian post-processing was applied using *BAYES* to improve the weak reflection data (Blessing *et al.*, 1996; French & Wilson, 1978). *LEVY* was used to estimate the absolute scale factor and the overall anisotropic mean-square displacement parameters and *EVAL* was used to generate normalized structure factors (Blessing *et al.*, 1996). Local scaling was applied using *LOCSCL* in an attempt to improve the SAS difference magnitudes (Blessing, 1997*a*; Matthews & Czerwinski, 1975) and, finally, *DIFFE* (Blessing & Smith, 1999) was used to produce the normalized difference structure-factor magnitudes or diff*E* values, denoted $|E_\Delta|$. The *DREAR* program suite can be run either as a stand-alone set of programs or by using the graphical interface provided in *SnB* version 2.0 (Weeks & Miller, 1999).

The program *BAYES* serves two purposes. Not only does it apply the Bayesian correction to the weakest data, but it also generates locally empirically normalized $|E|$ values. Typically, these $|E|$ values are only used for phasing in cases where the diffraction resolution limit does not extend beyond 3.5 Å, as under these circumstances the *LEVY* program is not applicable. Since the SAH1 and SAH2 data diffract to 2.8 and 2.6 Å resolution, respectively, all calculations described in this paper employed the globally Wilson-normalized $|E|$ values generated by the *LEVY* and *EVAL* programs. As described above, the program *BAYES* was used in this case to obtain improved estimates of $|F|$ and $\sigma(F)$ for the weak reflections.

2.2. Data-reduction parameters

The *DIFFE* program parameters $t_{\rm max}$, $x_{\rm min}$, $y_{\rm min}$, $z_{\rm min}$ and $z_{\rm max}$ are defined (Blessing & Smith, 1999) with respect to the following quantities:

$$\begin{aligned}
t &= |\Delta|/\sigma \\
&= \big|(|E_{+H}| - |E_{-H}|) - {\rm median}(|E_{+H}| - |E_{-H}|)\big|/ \\
&\quad [1.25\,{\rm median}\big|(|E_{+H}| - |E_{-H}|) \\
&\quad -\ {\rm median}(|E_{+H}| - |E_{-H}|)\big|], \\
x &= \min[|E_{+H}|/\sigma(|E_{+H}|), |E_{-H}|/\sigma(|E_{-H}|)], \\
y &= \big|\Delta|E|\big|/\sigma(\Delta|E|) \\
&= (\big||E_{+H}| - |E_{-H}|\big|)/[\sigma^2(|E_{+H}|) + \sigma^2(|E_{-H}|)]^{1/2}, \\
z &= |E_\Delta|/\sigma(|E_\Delta|) \text{ or } z = (|E_\Delta| - |E_\Delta|_{\rm max})/\sigma(|E|),
\end{aligned}$$

where $|E_\Delta|$ denotes a renormalized diff*E* value,

$$|E_\Delta| = [\textstyle\sum(f^0 + f')^2 + (f'')^2]^{1/2}\big||E_{+H}| - |E_{-H}|\big|/\{2q[\sum(f'')^2]^{1/2}\}$$

and $|E_\Delta|_{\rm max}$ is a physical least upper bound,

$$|E_\Delta|_{\rm max} = \textstyle\sum f''/[\varepsilon_H \sum(f'')^2]^{1/2}.$$

In the expression for $|E_\Delta|$ = diff*E*, the quantity

$$q = q_0 \exp(q_1 s^2 + q_2 s^4),$$

where s = $(\sin\theta_{\rm H})/\lambda$, is a least-squares-fitted empirical renormalization scaling function that imposes the condition $\langle|E_\Delta|^2\rangle$ = $\langle{\rm diff}E^2\rangle$ = 1 and serves to define q_0, q_1 and q_2.

Table 2
DREAR parameters used in tests and suggested default values.

Use of locally normalized $|E|$ values	No†
LOCSCL parameters	
Local scaling applied	No‡
$F/\sigma(F)$ cutoff	3
DIFFE parameters	
$t_{\max}$	6
$x_{\min}$	3
$y_{\min}$	1
$z_{\min}$	3
$z_{\max}$	0§

† Locally normalized $|E|$ values output from the *BAYES* program are not recommended for use unless the resolution of the data is less than 3.5 Å. ‡ Local scaling may be of benefit when the redundancy for the individual Bijvoet reflections is less than 5 or 6. § This parameter was not examined, since application of the default value resulted in no data being rejected.

The variable $t_{\max}$ is used to exclude data with unreliably large values of $\big||E_{+H}| - |E_{-H}|\big|$ in the tails of the $(|E_{+H}| - |E_{-H}|)$ distribution. This test assumes that the distribution of $(|E_{+H}| - |E_{-H}|)/\sigma(|E_{+H}| - |E_{-H}|)$ should approximate a zero-mean unit-variance normal distribution for which values of t less than $-t_{\max}$ or greater than $+t_{\max}$ are extremely improbable. The parameters $x_{\min}$ and $y_{\min}$ are a set of minimum criteria that must be met for any reflection pair to be included in the data processing, while $z_{\max}$ and $z_{\min}$ control the data that are output. The standard *DREAR* parameters used in the tests reported here, as well as the suggested default values, are listed in Table 2.

2.3. Standard *SnB* test

For each parameter examined (see Fig. 1), unless otherwise mentioned, a standard *SnB* test was run and examined for the number of solutions per 5000 trial structures (*i.e.* the success rate; see Table 3 for parameters used in the standard *SnB* test). The standard deviation of the success rate was calculated using the Bernoulli distribution estimate, $\sigma = (npq)^{1/2}$, where n is the number of trials, p is the success rate expressed as a fraction and q is the failure rate. The error bars drawn in the *SnB* success-rate plots indicate a significance of 1σ. The *DREAR* data-processing parameters were tested using the peak (λ_2) data for both SAH1 and SAH2, whereas tests to examine the parameters of the *SnB* phasing program were performed primarily using the SAH1 peak (λ_2) data.

Trial structures consisted of randomly positioned atoms that were refined by the dual-space *Shake-and-Bake* procedure as implemented in *SnB* v2.0. The phase-refinement portion of the dual-space cycle utilized parameter-shift reduction of the minimal function (Weeks *et al.*, 1994) and constraints were imposed in real space by density modification in the form of peak picking. Solutions were unequivocally identified on the basis of mean phase error by comparison with phases computed from the final refined coordinates of the Se atoms. The parameter values used for the standard tests are listed in Table 3. Many of these values are dependent on N_u, the number of unique non-H atoms in the asymmetric unit. In the case of selenium substructures, N_u is the number of independent Se atoms.

Table 3
Standard *SnB* parameters used for tests.

	Parameters used for tests†	Recommended values
Reflections and invariants		
Number of reflections	600	20–40N_u‡
Number of invariants	6000	200–400N_u
Resolution cutoff	None	None
Phase refinement		
Method	Parameter shift	Parameter shift
Phase shift (°)	90	90
Number of shifts	2	2
Number of passes through phase set	3	3
Fourier refinement menu		
(i) Real-space constraints		
Number of peaks to select	30	N_u
Fourier grid size	0.93 Å	1/3 $d_{\max}$
Minimum interpeak distance (Å)	2.8	3.0
Minimum distance between symmetry-related peaks (Å)	3.0	3.0
Number of special-position peaks to keep	0	0
(ii) Twice baking		
Fourier refinement	None	Best trial only
Number of cycles	3	0.1N_u
Number of peaks to select	30	N_u
Minimum $|E|$	0.75	0.75
Trials and cycles		
Starting phases from	Random atoms	Random atoms
Number of trials	5000	1000
Number of *SnB* cycles	60	2N_u

† The location of these parameters on the *SnB* graphical user interface (GUI) is shown in Fig. 1. ‡ N_u is the number of independent Se-atom positions.

3. Results and discussion

When determining a Se-atom substructure, a user will typically run only as many trials as are necessary to obtain a single solution or a small number of solutions. However, in this study each of the standard *SnB* tests involved 5000 trial structures, each of which was refined for 60 cycles. This protocol enabled the success rate associated with each parameter combination to be calculated accurately and the significance of any variation to be assessed in terms of the standard deviation. As described in §2, the rigorous identification of solutions required for the presentation of success-rate statistics was based on the computation of mean phase error. However, as has been observed for conventional full-structure applications of *SnB* (Miller *et al.*, 1994), *SnB* solutions for previously unknown structures can generally be clearly distinguished from non-solutions by the characteristic bimodal histogram of the minimal-function residual, $R_{\min}$ (data not shown). Fig. 2 illustrates the correlation between $R_{\min}$ values and the number of Se atoms that have been located. During some cycle, the value of $R_{\min}$ for a solution will drop significantly from its starting value, but the $R_{\min}$ values for non-solutions will remain high. If one examines the number of Se atoms found during each cycle, then the cycle where $R_{\min}$ suddenly decreases corresponds to the cycle where the number of Se atoms found dramatically increases. Thus, although the number of correct Se atoms found per cycle can only be determined in retrospect, solutions can also be easily identified from the characteristic step function in the $R_{\min}$ *versus* cycle number plot.

3.1. Data-processing (*DREAR*) parameters examined

3.1.1. Local scaling and minimum $F/\sigma(F)$ cutoff. Appropriate calculations were performed to determine whether or not local scaling of the data was advantageous and, if so, what effect a $F/\sigma(F)$ cutoff applied to the data used for local scaling might have. Values of 2 and 3 for the $F/\sigma(F)$ cutoff were investigated and found to have little effect on the *SnB* success rates for either SAH1 or SAH2 (data not shown). When the average $F^2/\sigma(F^2)$ values for both data sets are examined as a function of resolution (Table 1), it is obvious why the $F/\sigma(F)$ parameter has little effect. Even in the highest resolution shell, the average $F^2/\sigma(F^2)$ values are greater than 11.0 and 4.0 for SAH1 and SAH2, respectively.

The results presented in Fig. 3 show that there was no significant difference in the *SnB* success rate when local scaling was applied to the SAH2 data, but local scaling appeared to have a modest deleterious effect on the SAH1 success rates. The explanation of this counter-intuitive result might be that the SAH1 data were ‘too accurate’ for local scaling to be helpful. The local-scaling technique was developed almost 25 years ago (Matthews & Czerwinski, 1975) to deal with significant systematic errors such as those associated with inter-set scaling of data from multiple crystals or with time-dependent scaling to correct for crystal decay due to radiation damage in data measured at room temperature. Present-day cryo-crystallographic techniques largely obviate these problems and if, on average, the measurement error level is at or below the level of the anomalous difference signal, local scaling might be more hindrance than help since it might tend to diminish real differences between the measured $|F_{+H}|$ and $|F_{-H}|$. The anomalous SAH1 data were about sevenfold redundant and the measurements contained some 25% fully recorded reflections, whereas the SAH2 data were only about threefold redundant and contained no fully recorded reflections (Table 1). Thus, it is conceivable that local scaling slightly improved the less accurate SAH2 difference data, but actually degraded the more accurate SAH1 data.

3.1.2. Effect of f' and f'' values. At present, there are no experimental means to derive exact values of f' and f'' at most macromolecular synchrotron beamlines. Since these values depend on the local atomic environments of the anomalous scattering atoms and will vary from one protein to the next, values are typically estimated roughly and then refined against the MAD data. Since *DREAR* provides no means to refine the values of f' and f'', the effects that changes in these values might have on the final calculated difference E magnitudes (diff$E$s or E_{Δ}s) were investigated. The SAH2 peak data (λ_2) and three sets of f' and f'' values taken from the literature (set 1, −8.0161/7.88; set 2, −7.35/5.92; set 3, −7.64/4.43; Nagar *et al.*, 1998; Ramakrishnan & Biou, 1997; Doukov, T., personal communication) were used to calculate diff$E$s. Comparison of the diff$E$s calculated for each of these f' and f'' values found that the diff$E$ value for individual reflections typically differed by less than 0.001 (data not shown). Although some differences were as much as 0.01, such cases involved only a small percentage of the data (∼2–4%) and represented a change in the diff$E$ value of less than 1%. The effect of changing the f' and f'' values is absorbed into the diff$E$ renormalization scaling function. This function contains a polynomial (see §2.2), the coefficients of which are calculated such that $\langle \mathrm{diff}E^2\rangle = 1$. Modest differences in f' and f'' values will be absorbed into the coefficients q_0, q_1 and q_2 and will therefore make very little difference to the magnitudes of the diff$E$s calculated or, consequently, the success rates achieved in *SnB*. This result does not imply that accurate f' and f'' values are not critical for other parts of the structure-solution process.

3.1.3. Diff$E$ parameters. The *DIFFE* program parameters (Blessing & Smith, 1999) $t_{\max,}$ $x_{\min}$, $y_{\min}$, $z_{\min}$ and $z_{\max}$ (see §2 for the definition of these parameters) have also been examined. For the $z_{\max}$ parameter, it was found that even for $z_{\max} = 0$, no reflections were rejected; therefore, the effect of this parameter could not be determined. The $t_{\max}$ parameter

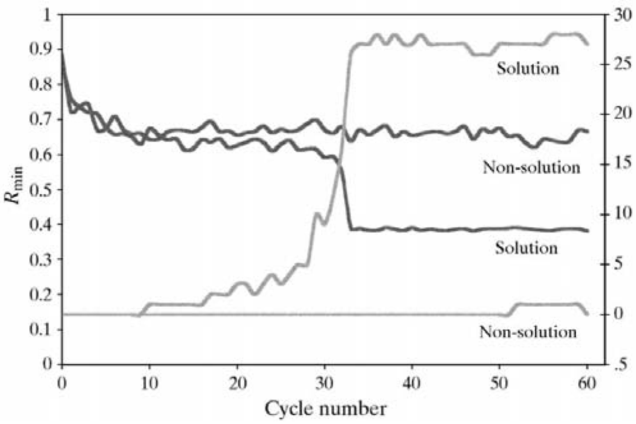


Figure 2
Plot of $R_{\min}$ (black line) and the number of correct Se-atom positions found (gray line) *versus* cycle number for a solution and a non-solution.

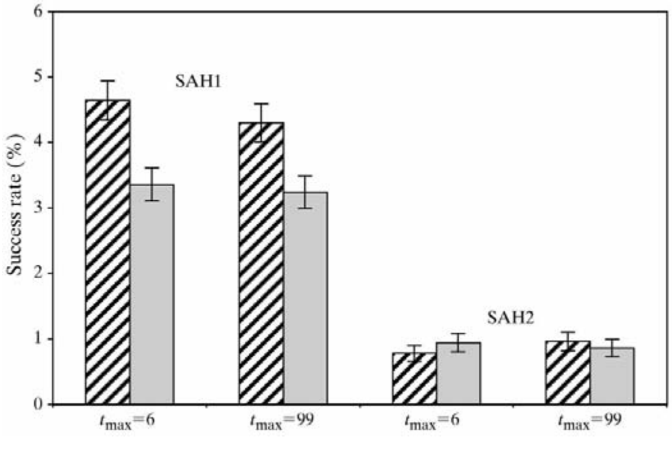


Figure 3
Plot of *SnB* success rates for SAH1 and SAH2 examining the effect of local scaling and the value of the *DIFFE* parameter $t_{\max}$. Solid gray or black cross-hatching represents data that were or were not locally scaled, respectively. Error bars are drawn at ±1σ.

was examined at the same time that the effect of local scaling was investigated and the results are presented in Fig. 3. For SAH1, removal of data with unreliably large $\left||E_{+H}| - |E_{-H}|\right|$ values ($t_{\rm max}$ = 6) appears to have a small but not statistically significant effect on the overall success rate. The effect of $t_{\rm max}$ on the SAH2 data, while again not statistically significant, appears to depend on whether the data are locally scaled or not. When local scaling is applied, culling out the large differences appears to be a slight improvement over no truncation of the data ($t_{\rm max}$ = 99).

Since the standard *SnB* calculation (Table 3) takes 13.9 CPU days on an R10K processor, an exhaustive comparison of all possible combinations of data-processing parameters was not possible with the computing facilities available. Given the results presented in Fig. 3, an $F/\sigma(F)$ cutoff of 2 with no local scaling and $t_{\rm max}$ = 6 were used as the default settings when examining the $x_{\rm min}$, $y_{\rm min}$ and $z_{\rm min}$ parameters (Figs. 4*a*–4*f*). As previously observed in other circumstances, the rates of success for the different values of $x_{\rm min}$, $y_{\rm min}$ and $z_{\rm min}$ depend on whether one examines the SAH1 or SAH2 data. For example, the *SnB* success rates for the SAH2 data appear to be largely insensitive to the values of the diff*E* parameters (Figs. 4*d*–4*f*). In contrast, the SAH1 data exhibit a maximum success rate when $x_{\rm min}$ and $z_{\rm min}$ are each equal to 3. When $x_{\rm min}$ and $z_{\rm min}$ are fixed and $y_{\rm min}$ is varied, there appears to be a steady increase in the success rate, with values of $y_{\rm min}$ > 1.0 having the best success rate. The trend in $z_{\rm min}$ is probably related to the number of diff*E*s retained for use in *SnB* (Fig. 4*c*). When $z_{\rm min}$ is too low, the number of reflections that are included is large, but the significance and quality of much of these data [diff*E*/σ(diff*E*)] are suspect, causing the success rate to decrease. Conversely, when $z_{\rm min}$ is too high, reflections that have accurate diff*E* values and make a significant contribution to the *SnB* phasing procedure are discarded, again lowering the success rate. A $z_{\rm min}$ value of 2.5 or 3 appears to be optimum.

While the parameter $z_{\rm min}$ controls the number of diff*E*s that are output, $x_{\rm min}$ and $y_{\rm min}$ control the number of pairs of reflections that are input into the renormalization procedure. The parameter $x_{\rm min}$ rejects reflections based on the significance of individual pairs of reflections [$|E_{+H}|/\sigma(|E_{+H}|)$, $|E_{-H}|/\sigma(|E_{-H}|)$] and, in the present case, appears to have little effect on the overall number of reflection pairs excluded (see Fig. 4*a*). The difference in the number of SAH1 reflection pairs rejected between $x_{\rm min}$ = 1.0 and $x_{\rm min}$ = 5.0 is only about 2000 in 26 000 reflections. The parameter $x_{\rm min}$ behaves like $z_{\rm min}$ in that an intermediate cutoff value appears to be best. The parameter $y_{\rm min}$ rejects reflections on the basis of the significance of the anomalous difference $\{\left||E_{+H}| - |E_{-H}|\right|/[\sigma^2(|E_{+H}|) + \sigma^2(|E_{-H}|)]^{1/2}\}$ and, unlike $x_{\rm min}$, it has a substantial influence on the

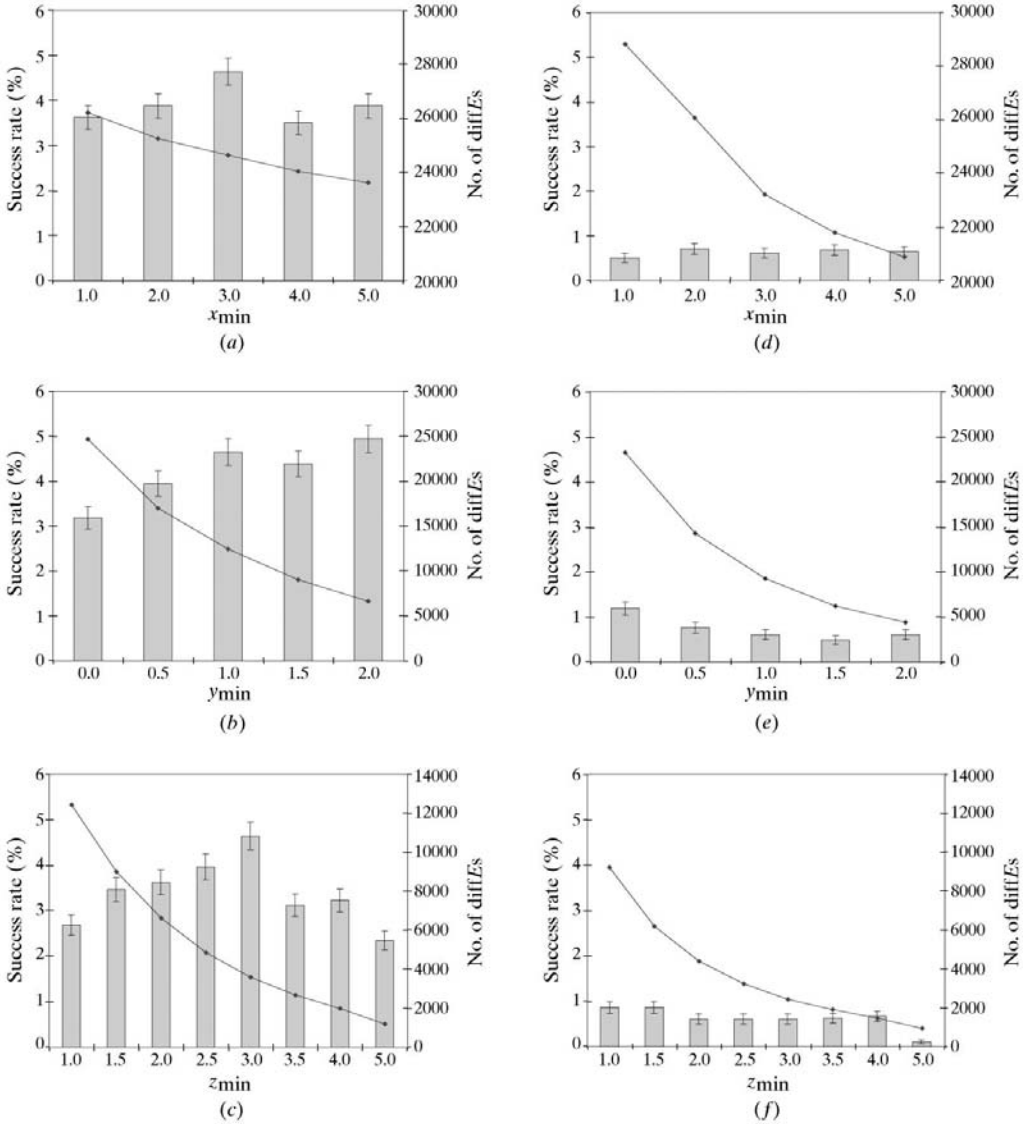


Figure 4

Histogram of *SnB* success rates for SAH1 (*a*–*c*) and SAH2 (*d*–*f*) examining the effect of the *DIFFE* parameters $x_{\rm min}$ (*a*, *d*), $y_{\rm min}$ (*b*, *e*) and $z_{\rm min}$ (*c*, *f*). For $x_{\rm min}$ and $y_{\rm min}$, the numbers of reflection pairs included in the data processing are plotted on a second axis; for $z_{\rm min}$, the total number of resulting diff*E*s are plotted. When not being varied, $x_{\rm min}$, $y_{\rm min}$ and $z_{\rm min}$ were held fixed at their default values of 3, 1 and 3, respectively. Error bars are drawn at ±1σ.

Table 4
Effect of atom:phase and phase:invariant ratios on *SnB* success.

Atom: phase ratio	Phase:invariant ratio				
	1:5	1:10	1:20	1:30	1:40
1:10	2.80 ± 0.23	2.46 ± 0.22	isp†	isp	isp
1:20	**5.34 ± 0.32**	**4.38 ± 0.29**	3.22 ± 0.25	isp	isp
1:30	**4.10 ± 0.28**	**4.08 ± 0.28**	3.26 ± 0.25	2.74 ± 0.23	2.48 ± 0.22
1:40	3.74 ± 0.27	3.46 ± 0.26	2.66 ± 0.23	2.54 ± 0.22	1.82 ± 0.19
1:50	3.74 ± 0.27	3.48 ± 0.26	nd‡	nd	nd
1:60	3.54 ± 0.26	3.08 ± 0.24	nd	nd	nd

† isp = insufficient phases to calculate the required number of triplet-phase invariants. ‡ nd = not determined.

number of reflection pairs excluded (see Fig. 4*b*). It seems likely that optimum choices for these parameters are, in fact, dependent on each other to some degree, with the default values being a well balanced combination. The fundamental importance of rejecting the less reliable anomalous differences is, however, emphasized by an experiment in which no rejection criteria were applied to the SAH1 data (*i.e.* $x_{\min} = y_{\min} = z_{\min} = 0$). In this case, the success rate for the standard test was 1.26 ± 0.01%. The reflection file input into *SnB* included nine high-resolution reflections with diff*E* values in the range 5.16–8.82 as well as a number of additional reflections (reflections that are normally excluded by the cutoffs) having diff*E* values greater than 4.0. The *SnB* parameter $|E_{\max}|$ is a last-resort mechanism for eliminating reflections with unrealistically high $|E_{\Delta}|$ values. Setting $|E_{\max}|$ to 5.0 eliminated the nine worst reflections and resulted in a success rate of 1.62 ± 0.01%. However, this is still only about one third of the best success rate obtained with the default cutoffs.

3.1.4. Use of edge and remote data. An important advantage of using the diff*E* method in conjunction with *SnB* is that the edge (λ_1) and remote (λ_3) data can be treated independently of the peak data. To investigate whether the edge and remote data could be used to determine the Se-atom substructure, the standard *SnB* test was applied to these data for both SAH1 and SAH2. The results presented in Fig. 5 clearly show that all three wavelengths of MAD data were capable of locating the Se atoms. In the case of SAH1, the *SnB* success rates for the edge and remote data were much lower than that of the peak data. In contrast, the SAH2 remote data were found to have the highest success rate. The reason for this unexpected result is unclear. The high success rate for the remote SAH2 data suggests that if there are difficulties solving the substructure with the peak data, then the other wavelengths should be examined. In addition, the ability to use each wavelength separately permits independent verification of the Se-atom positions (discussed in greater detail below) and this can be useful even when the peak data seem to give routine solutions.

3.2. *SnB* parameters examined

3.2.1. Number of phases and triplet-phase invariants. The effects of the number of phases (or reflections) and triplet-phase invariants (linear combinations of phases whose Miller indices sum to zero) used in the *SnB* phasing process have been examined and the results for SAH1 are presented in Table 4. Maximum success rates occur with atom:phase ratios in the range 1:20 to 1:30 and phase:invariant ratios in the range 1:5 to 1:10. As the numbers of phases and invariants are increased further, the *SnB* success rates decrease. This observation is consistent with the fact that the accuracy with which the value of a triplet-phase invariant can be estimated decreases as the product of the three associated $|E|$ values decreases (Cochran, 1995). Thus, it is standard practice in small-molecule direct-method applications to limit the number of phases to about ten times the number of atoms sought and to use those phases corresponding to the largest E magnitudes, which typically have minimum $|E|$ values in the range 1.3–1.5. Experience has shown that this is an adequate number of terms to include in the Fourier summation for computation of an electron-density map. In *SnB*, the invariants are sorted in accordance with the triple product of $|E|$ values and only the requested number of most reliable invariants are used in the phasing process.

Table 4 also shows that when the number of reflections is small (*i.e.* atom:phase ratio less than 1:20), the number of triplet-phase invariants that can be generated is greatly reduced and the numbers of invariants available may be insufficient to satisfy the larger phase:invariant ratios (*i.e.* greater than 1:20). In general, a given number of reflections with the largest $|E|$ values generate fewer invariants for a substructure than they would for a real small-molecule structure having the same number of independent atoms. This happens because the substructure cell is much larger and there is a reduced probability that the indices of any three large $|E|$ reflections will sum to zero. For this reason, the default atom:phase ratio for substructures is 1:30, but the phase: invariant ratio (1:10) is the same as for small molecules. If *SnB* fails to generate a suitable combination of phases and invariants automatically, then adjustment of these parameters (more phases or fewer invariants) is required.

3.2.2. Real-space constraints. Parameters that affect the way in which the electron density is modified during the real-

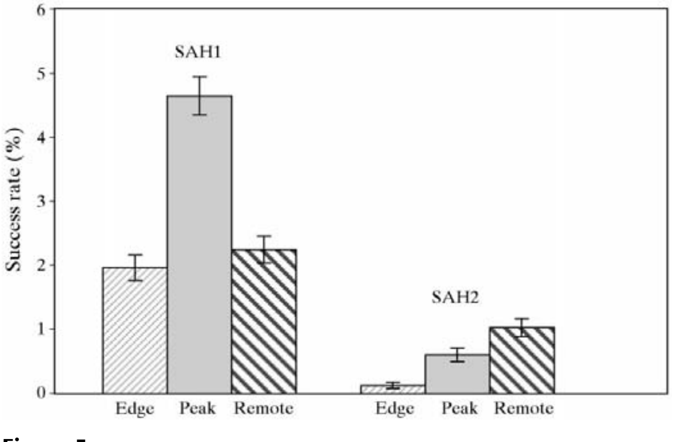


Figure 5
Plot of *SnB* success rates for diff*E*s calculated from individual edge, peak and remote data sets for SAH1 and SAH2. Error bars are drawn at ±1σ.

space portion of the dual-space *Shake-and-Bake* cycle apply powerful constraints to the trial phases. The peak-picking algorithm used in *SnB* v2.0 has three such parameters. They control where in the unit-cell peaks are permitted to be, the geometrical restrictions on interpeak distances and the total number of peaks that can be selected in each cycle for use as atoms in the subsequent structure-factor calculation.

Since it is impossible for Se atoms substituting for methionine S atoms to lie on any special positions that exist in a protein crystal structure, this information can be used during selenium-substructure determination. *SnB* v2.0 has a parameter that specifies the minimum distance permitted between symmetry-related peaks; peaks that violate this condition are eliminated automatically. In space group *C*222, this distance parameter specifies the diameter of a cylinder of excluded volume about each rotation axis. Minimum distances between symmetry-related peaks of 0, 3 and 6 Å were tested and the *SnB* success rates were found to be 0.8 ± 0.13, 3.28 ± 0.25 and 3.52 ± 0.26%, respectively. Thus, cylinders of either 3 or 6 Å give a highly significant improvement in success rate. In fact, the difference is even more dramatic than is shown by these numbers because failure to exclude disallowed peaks (*i.e.* an *SnB* calculation with 0 Å distance) permits, in a few cases, the development of false minima that cannot be distinguished from solutions based on their $R_{\min}$ values. The use of this distance parameter as well as other methods for avoiding false minima in *SnB* have recently been described in detail (Xu *et al.*, 2000). The differences in success rates using the 3 and 6 Å minimum distances are not significant and a distance close to the van der Waals distance of the atoms involved is probably a good choice. The supplied default distance of 3 Å should be adequate for most Se-atom substructures.

The second real-space constraint parameter is the minimum permitted interpeak distance. The results of varying this distance from 0.93 to 5.6 Å are presented in Fig. 6(*a*). This parameter, which excludes peaks if they fall closer than the chosen value to a larger peak, appears to have little effect on the *SnB* success rate for AdoHcy hydrolase. The largest distance (5.6 Å) gives the highest success, but there is no consistent trend as the distance decreases. This result is not surprising as only two selenium–selenium distances in this structure are less than 5.6 Å. This selenium–selenium distance distribution appears to be typical of protein structures in general. Examination of 100 randomly chosen structures from the Protein Data Bank with a total of 549 methionine-sulfur-to-methionine-sulfur distances less than 10 Å revealed that the shortest sulfur–sulfur distance was 2.99 Å and only 15 such distances were less than 4 Å. Since the minimum selenium–selenium distance cannot be predicted *a priori*, it seems reasonable that the minimum interpeak distance be set to a value approximating twice the van der Waals radii for an Se atom. A default value of 3 Å is suggested.

The third real-space constraint parameter is the number of peaks to be selected in each cycle. This value was varied from 15 to 50 peaks. Fig. 6(*b*) shows that there was no statistically significant variation in the success rate, although there is some indication that selecting too many peaks is not favorable. By default, *SnB* selects N_u peaks (*i.e.* the number of expected sites). In some cases, however, it may be preferable to pick fewer peaks (*e.g.* $0.9N_u$), since some of the sites may be disordered or have high thermal motion and therefore be unlikely to show up in preliminary maps. This is especially true for structures having a very large number of sites.

Inherently linked with the number of peaks selected are the important questions of which peaks and how many peaks from the output list actually correspond to atomic positions. This list is sorted in decreasing order according to peak height. Although $1.5N_u$ peaks are available for consideration, the first (largest) N_u peaks are most likely to be correct. Solutions obtained from all three wavelengths of data for both SAH1 and SAH2 were examined to determine how many peaks among the top 30 were correct; the results of this study are presented in Fig. 7. Typically, 28–30 of the 30 Se atoms were found for both sets of peak and remote data. Fewer peaks were located for the edge data, especially for SAH1. The reason for this is unclear.

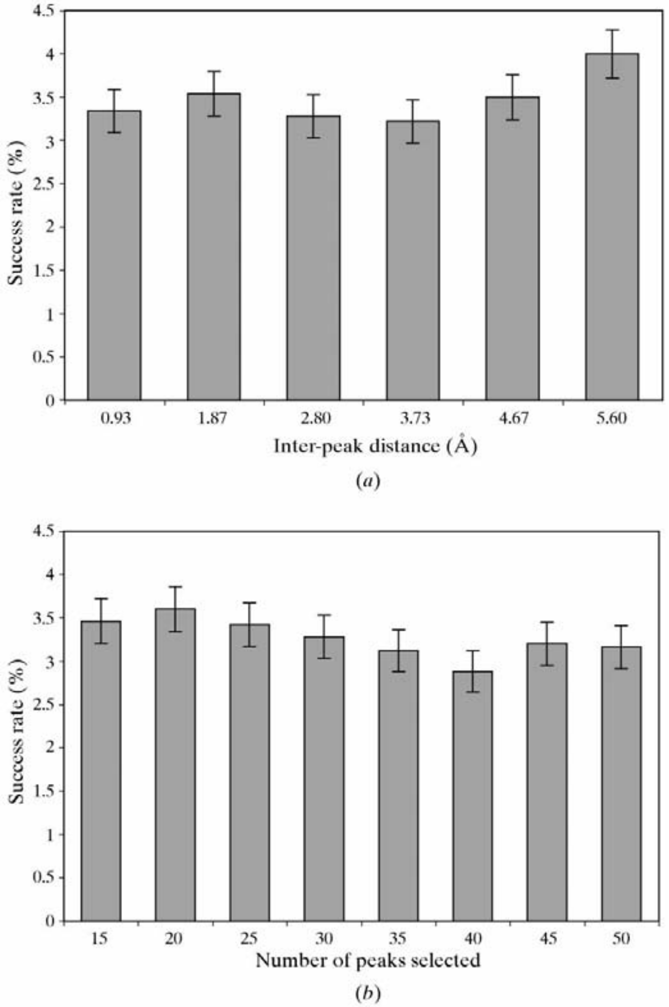


Figure 6
Plot of *SnB* success rates for (*a*) the minimum interpeak distance permitted and (*b*) the number of peaks selected as Se atoms in each cycle. Error bars are drawn at ±1σ.

Table 5
Peak correlation and Fourier refinement results for ten solutions.

Crystal	SAH1							SAH2		
Wavelength	Peak					Edge	Remote	Edge	Peak	Remote
Trial number	149	31	158	165	176	104	23	476	93	86
(*a*) Peak rankings before Fourier refinement										
	16	19	16	17	14	19	31	19	16	15
	21	25	23	21	24	35	29	17	19	1
	27	27	27	27	27		15	39	21	8
	29	22	29	29	29		21	38	29	28
	31			34						
	33	42	30	30	35	24	22	34	30	30
	34		33		42					
	37			43						
	39		40		38					
	40		42			42				
	45	40								
(*b*) Map improvement following Fourier refinement										
Before refinement										
Atoms found	29	29	30	30	29	24	29	27	30	30
⟨Distance⟩†	0.32	0.31	0.33	0.31	0.31	0.32	0.38	0.42	0.36	0.48
After refinement										
Atoms found	30	30	30	30	30	25	30	28	30	30
⟨Distance⟩†	0.28	0.30	0.28	0.29	0.30	0.30	0.35	0.33	0.33	0.40

† Average deviation (Å) from the corresponding refined Se-atom positions.

When trying to decide which peaks are correct, it is helpful to compare the peak positions from two or more *SnB* solutions and to determine which peaks they have in common. Peaks occurring in several solutions (especially solutions obtained from data measured with different wavelengths) are more likely to be real. However, in order to perform this comparison, it is necessary to take into account the fact that different solutions may have different origins and/or enantiomorphs. A stand-alone program for doing this is available from one of the authors (GDS); the capability of making such comparisons automatically for all space groups will be available in a future version of *SnB*. The usefulness of peak correlation is illustrated by the example of the ten solutions given in Table 5(*a*), which presents the relative rankings of corresponding peaks on the different maps. The top 29 peaks were correct Se-atom positions for trial 149 for the SAH1 peak data. Peak 30 was spurious. Table 5 lists peaks from other trials that match peaks 1–29 of trial 149 if their peak height ranking exceeds 31 and also lists peaks of trial 149 that were not found in other trials [*e.g.* trial 104 (SAH1 edge data) failed to find peaks 27 and 29]. Rankings are also listed for all peaks matching peaks 30–45 of trial 149. Peak 33 of trial 149 was found to have a match on every other map and indeed also corresponds to the missing 30th Se atom. Thus, peak correlation can be used to identify correct peaks ranking above peak 30.

3.2.3. Fourier refinement. Fourier refinement, often called E-Fourier recycling, has been used for many years in conventional direct-method programs to improve the completeness of solutions following tangent-formula phase refinement (Sheldrick, 1982). Similarly, the 'twice-baking' option allows individual *SnB* trials to be refined for additional cycles in real space alone. The parameters of this procedure and their default values, as determined for full structures in the 300–500 atom range, are as follows: number of cycles ($0.1N_u$), number of peaks (N_u) and minimum $|E|$ value (0.75). Since invariant accuracy is not a concern during Fourier refinement, additional reflections are normally included to reduce series-termination errors. In *SnB*, these additional reflections are gradually introduced in equal lots during the specified number of cycles. In order to save computing time, Fourier refinement is typically applied to a trial structure only if it is identified, by virtue of its $R_{\min}$ value, as the current best trial.

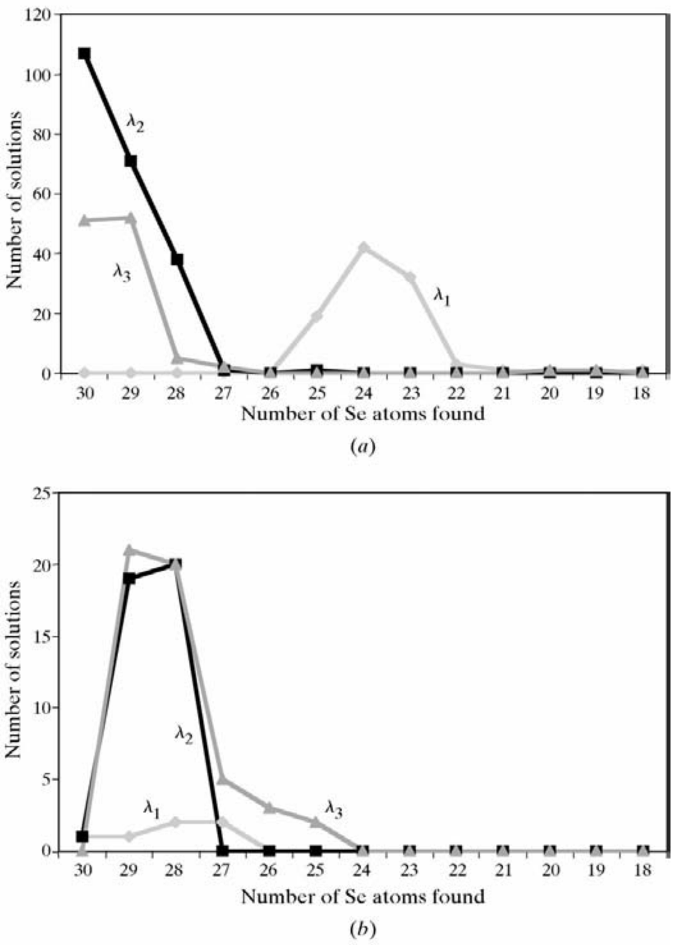
AdoHcy hydrolase is not a good test case for judging the value of Fourier refinement for substructures in general, since all, or almost all, of the Se atoms are typically found during the dual-space refinement cycles. Nevertheless, the trials used to investigate peak correlation were refined in this way to see if any improvement could be obtained. The results, presented in Table 5(*b*), show that use of the default refinement conditions locates an additional atom in all cases where atoms were originally missing. Fourier refinement has also improved slightly the average deviation between the peaks and the corresponding final refined Se-atom positions. This improvement was most noticeable for those trials having the largest deviations prior to Fourier refinement. The effects of such refinement may be more significant for larger substructures for which a smaller percentage of correct Se atoms were initially found.

3.2.4. Data resolution and grid size. In order to gain some understanding of the data resolution required for *SnB* applications, the 2.8 Å SAH1 peak data were truncated to 4 and 5 Å. Fig. 8(*a*) shows that both of these truncated data sets were capable of yielding solutions. The percentage success rate for the truncated 5 Å data was, however, significantly less than that for either the complete 2.8 Å or the truncated 4 Å data. On the other hand, success rate is not an adequate measure of computational efficiency. This needs to be accessed in terms of the number of solutions obtained per unit of CPU time, a quantity that has been called the *cost effectiveness* of the *SnB* procedure (Chang *et al.*, 1997). In this case, consideration of the number of solutions found per hour (Fig. 8*b*) shows that both truncated data sets generate solutions more efficiently than the 2.8 Å data. Thus, if the computational overhead of the *SnB* procedure for a particular substructure is a concern, then the data can be truncated and solutions might be generated more rapidly. However, in such cases the total number of Se atoms found may be compromised. For example, the truncated 5 Å SAH1 data typically found only 25 of the 30 Se atoms. Correction solutions can however always be recalculated as a single trial run using higher resolution data, if available, in order to increase the number of Se atoms found. Nevertheless, it is clear that successful *SnB* substructure applications should be possible for data sets having a resolution less than 3 Å. This prediction was confirmed by the recent solution of a 48-selenium substructure using only 4 Å data (Noble, personal communication).

It is common practice to compute Fourier maps using a grid spacing approximately equal to 1/3 of the data resolution and the default *SnB* grid is chosen in this way. However, substructure cells contain mostly empty space, so the SAH1 peak data were used to investigate whether a coarser grid would be sufficient. The data presented in Table 6 indicate that coarser grids can indeed provide solutions more efficiently than the standard 0.93 Å (1/3 resolution) grid because the Fourier summation needs to be computed at many fewer points. These solutions are recognizable on the basis of minimal function values ($R_{\rm min}$), but the complete substructure is not found. This can be easily fixed by using the largest peaks in the output file as the starting point for a second (single-trial) run with a standard-size grid. Although this two-step process currently requires user intervention, it may save a great deal of computing time for large substructures.

3.2.5. Trials and cycles. The default number of trials is 1000. In most cases, this many trials will not be necessary and the user can terminate the calculation after examining the $R_{\rm min}$ histogram and discovering a clear bimodal distribution that indicates that solutions are present. A good strategy is to permit the program to find a small number of probable solutions and then compare them as described above in order to define a common set of atomic positions. Very large substructures may require more trials. Failure to find any likely solutions after several thousand trials have been examined suggests that there may be a problem (*e.g.* an inadequate anomalous signal) with the data.

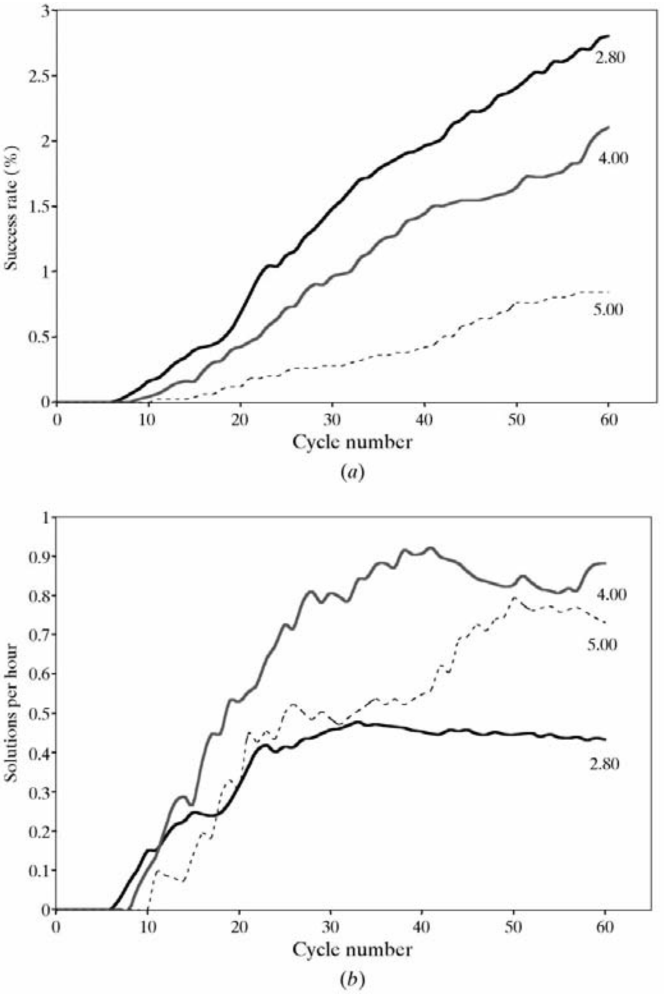
The choice of number of dual-space *Shake-and-Bake* cycles can be critical. Although increasing the number of cycles per *SnB* trial increases the computational overhead, performing too few cycles can result in failure to find any solutions at all.



(*a*)

(*b*)

Figure 7
Plots of the number of correct Se-atom positions found *versus* the number of solutions found in the 5000 trials performed for (*a*) SAH1 and (*b*) SAH2. The results for the edge, peak and remote (λ_1, λ_2, λ_3) data are presented in light gray, black and medium gray, respectively.



(*a*)

(*b*)

Figure 8
Effect of truncating the resolution of the data on the (*a*) *SnB* success rates and (*b*) solutions per hour or cost effectiveness for SAH1 peak data. Each graph is plotted against the *SnB* cycle number.

Table 6
Effects of varying the Fourier map grid size.

Grid size (Å)	Success rate (%) ± s.d.	Solutions per hour	Highest $R_{\min}$ for a solution	Lowest $R_{\min}$ for a non-solution	Best trial: mean phase error (°)	Best trial: No. atoms found
0.93	4.64 ± 0.30	0.8	0.36	0.37	5.8	30
1.5	2.00 ± 0.20	0.9	0.40	0.54	12.0	25
2.0	1.12 ± 0.15	2.4	0.45	0.54	20.6	12
2.5	1.02 ± 0.14	3.4	0.50	0.57	26.6	10
3.0	0.32 ± 0.08	1.5	0.57	0.63	38.8	6

Fig. 2 shows the characteristic plot of $R_{\min}$ *versus* cycle number for both a solution and a non-solution. In this case, the solution starts to emerge (indicated by the sharp decrease in $R_{\min}$) at cycle 33. If fewer cycles had been performed, this solution would have been lost.

The optimum number of cycles can be determined by considering the efficiency of the *SnB* procedure as a function of cycle number. Under conditions where the user is monitoring the progress of his calculation, the least CPU time required to achieve a solution will be when the most cost-effective number of cycles is chosen. Fig. 9 illustrates cost effectiveness for the following cases: (*a*) 2.8 Å SAH1 peak data, (*b*) truncated 5 Å SAH1 peak data, (*c*) 2.6 Å SAH2 peak data and (*d*) 2.6 Å SAH2 remote data. Top efficiency is reached by N_u cycles for both sets of peak data at high resolution, but more cycles are required for the truncated SAH1 data or the SAH2 remote data. Based on the assumption that it is better to choose defaults that optimize the least ideal situations, a default value of $2N_u$ cycles has been selected for substructure applications. When the number of Se atoms to be found is less that ten, it is recommended that at least 20 cycles be run for every trial in order to minimize the chance of failure in situations where the amount of computing time required is not large anyway.

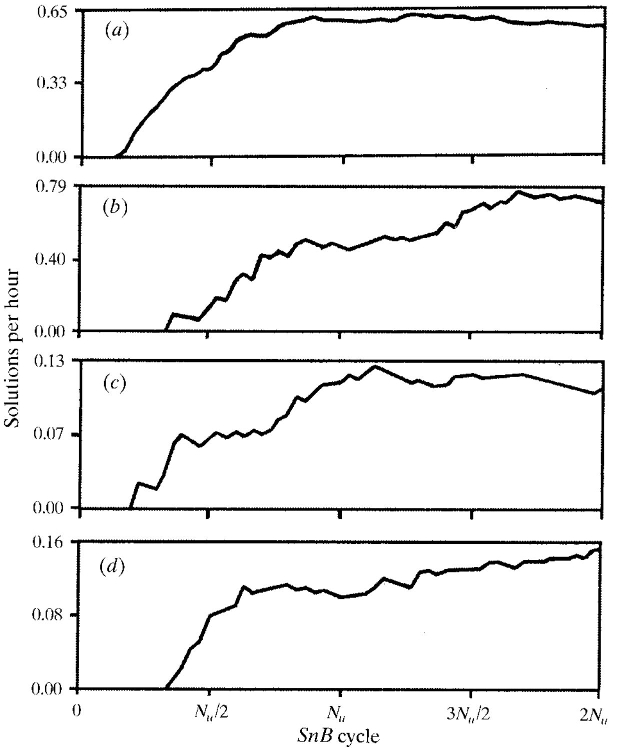


Figure 9
Cost effectiveness as a function of cycle number for the following cases: (*a*) 2.8 Å SAH1 peak data, (*b*) 5 Å SAH1 peak data, (*c*) 2.6 Å SAH2 peak data and (*d*) 2.6 Å SAH2 remote data.

3.3. Examination of the differences between SAH1 and SAH2 data

Although both the SAH1 and SAH2 data successfully determine the Se-atom substructure, there is a considerable (approximately fourfold) difference in the *SnB* success rate. Therefore, an investigation was undertaken to try to identify the cause or causes of this difference and to learn how to optimize MAD data collection in the future. One obvious difference between the data for the two crystals is the amount of low-resolution data that were measured. As noted in Table 1, SAH1 contains data between 50 and 2.8 Å resolution, while SAH2 contains data in the range 20–2.6 Å. Since the crystal-to-detector distance was comparable (*i.e.* 137 mm *versus* 142 mm for SAH1 and SAH2, respectively), the difference in minimum data resolution was the result of a difference in the position of the backstop. It is known that low-resolution data can play an important role in direct-methods procedures because such reflections tend to form large numbers of three-phase structure invariants. Therefore, the SAH1 data were truncated to 20 Å and a standard *SnB* test was run. What is striking about the results presented in Fig. 10 (light gray cross-hatching) is that truncation of the SAH1 data results in a greater than 50% reduction of *SnB* success rate. This is significant considering that only six reflections (222, 331, 113, 333, 171 and 242) were eliminated from the set of diff*E*s actually used for *SnB* phasing.

Since the truncated 20 Å SAH1 data still have a success rate nearly twice that of the SAH2 data, other factors must also be important. Several relevant properties of the 600 largest diff*E*s used in the standard tests are compared in Table 7. Using the refined selenium coordinates, thermal factors and occupancies, $|E_{\rm calc}|$ values were generated and compared to the diff*E* values from *DREAR*. Computation of the differences, $\big|{\rm diff}E - |E_{\rm calc}|\big|$, showed that outliers occurred with approximately equal frequency in the two data sets and the overall crystallographic *R* values were the same. On the other hand, examination of the minimum $|E|$ values revealed that SAH1 has more very large diff*E* magnitudes, especially at resolutions higher than 4 Å (Fig. 11). It follows, therefore, that SAH1 also has structure-invariant relationships with higher *A* values and, presumably, better estimated triplet-phase values. In addition, normal probability-plot analysis (Howell & Smith, 1992) showed that the SAH2 diff$E$ − $|E_{\rm calc}|$ differences deviated

Table 7
Comparison of the 600 SAH1 and SAH2 reflections used in the standard *SnB* tests.

	SAH1	SAH2
Number of $\lvert \mathrm{diff}E - \lvert E_{\mathrm{calc}}\rvert\rvert > 1.0$	16	16
Max. $\lvert \mathrm{diff}E - \lvert E_{\mathrm{calc}}\rvert\rvert$	1.42	1.58
R factor = $\sum\lvert \mathrm{diff}E - \lvert E_{\mathrm{calc}}\rvert\rvert/\sum \mathrm{diff}E$	0.21	0.21
Diff$E$ minimum	1.68	1.53
Invariant A values		
Maximum	8.41	4.47
Minimum	1.48	1.33
Average	1.91	1.69
Normal probability plot residual	0.13	0.22

more from a normal distribution of errors than did the SAH1 differences, especially at low resolution.

It is also interesting to note that the success rate for the 600 largest $|E_{\mathrm{calc}}|$ values in the standard SnB tests was 3.3% or less than that of the SAH1 data. Removing the zonal (centric) reflections from the calculated data set improved the success rate to 3.8%, possibly because there are more reliable invariant relationships involving the general reflections alone. These observations illustrate how, in substructure problems where the percentage of data actually used for phasing is small (approximately 1/50 of the total possible reflections compared with approximately 1/6 of the data in the typical small-molecule structure), considerable variation in success rate is possible depending on the exact identities of the reflections involved.

To investigate further how the choice of particular phases influenced the *SnB* success rate, the largest 600 diff*E*s from SAH1 that were common to both the SAH1 and SAH2 data were chosen for a test run. Since identical reflections (*i.e.* identical *hkl*s) were used for both the SAH1 and SAH2 data, this test addressed the question of how the magnitudes of the diff*E*s, which are directly related to measurement accuracy, influenced the success rate. The results presented in Fig. 10 (columns with solid gray shading) show that the *SnB* success rate for SAH1 still appears to be significantly different from, and superior to, that found for the SAH2 data. The top 600 common reflections for SAH2 also appear to have a higher success rate than the regular SAH2 data, although this difference is not significant. The identities of 165 reflections differ between the largest 600 diff*E*s in the complete SAH2 data and the 600 diff*E*s common to the SAH1 data. Examination of these 165 data with regards to resolution and diff*E* magnitude revealed no obvious trends. One explanation for the observed success rates is that selecting SAH2 diff*E*s that are large in SAH1 eliminates 165 small diff*E*s that are erroneously estimated to be large in SAH2. This result supports the general hypothesis that the quality of the SAH2 amplitudes is inferior to those of SAH1 and that this difference can be attributed to the fact that the SAH1 data are more redundant than SAH2 data (*i.e.* average redundancies of 7 and 3, respectively, for anomalous pairs) and are therefore likely to be more accurate.

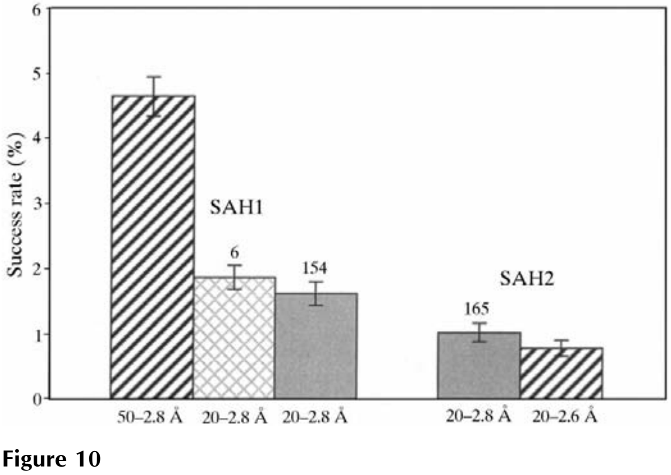


Figure 10
Plot of *SnB* success rates for SAH1 and SAH2 for the standard *SnB* test (black stripes) and after truncation of the low-resolution SAH1 data. When only data in the range 20–2.8 Å resolution (light gray cross hatching) were included, six reflections were replaced among the top 600 SAH1 diff*E*s used for *SnB* phasing. Next, the truncated SAH1 and the SAH2 diff*E*s were compared and the first 600 reflections common to both data sets were retained for additional test runs (solid gray). When these SAH1 and SAH2 data were compared to the original data, 154 (SAH1) and 165 (SAH2) diff*E*s were found to differ from their parent data sets. Error bars are drawn at ±1σ.

To examine the hypothesis that more highly redundant data yield higher *SnB* success rates, a series of standard *SnB* tests were run with the SAH1 data truncated to be either threefold or fivefold redundant. To mimic a typical X-ray diffraction experiment, contiguous frames of the SAH1 data were scaled together. Since each frame is a 1° oscillation and inverse-beam geometry was applied, a total of 180 or 260 frames were used for the threefold and fivefold redundant data sets, respectively. The corresponding R_{merge} values were 6.8 and 7.0%, respectively, and the corresponding overall completenesses were 95 and 98%, respectively. These values are comparable to those found for the SAH1 and SAH2 data (see Table 1). The SAH1 and SAH2 data are sevenfold and threefold redundant, respectively. Fig. 12 shows clearly that the *SnB* success rate decreases when the redundancy of the SAH1 data is reduced. The *SnB* success rates for the threefold-redundant data are

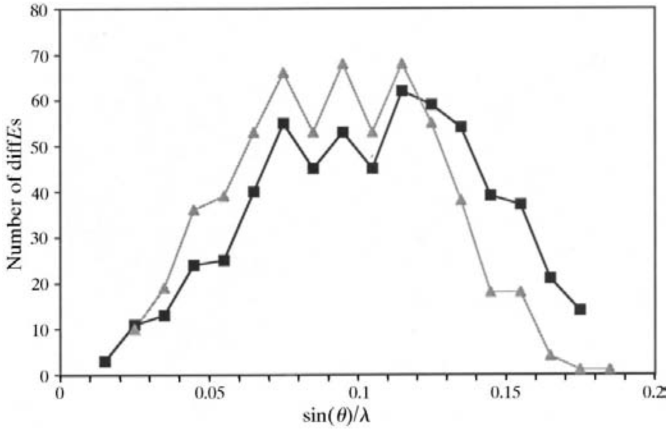


Figure 11
Distribution of diff*E*s as a function of sinθ/λ for SAH1 (black line) and SAH2 (gray line) data.

four to five times lower than those for the sevenfold-redundant data, with the success rates for the fivefold-redundant data being intermediate between the other two. It is also interesting to note that the threefold-redundant SAH1 data give comparable success rates to those found for the SAH2 data, an observation which supports the hypothesis that the quality of the data is directly related to redundancy and that the higher the redundancy, the better the *SnB* success rate.

4. Conclusions

Successful application of the *SnB* program to substructure difference data requires careful data collection, processing and computation of normalized structure-factor difference magnitudes. It is essential that all erroneous and unrealistically large magnitudes be identified and excluded before phasing is attempted. This study has highlighted not only the necessity for accurate high-resolution data, but also the importance of the very low-resolution data. It is worth the time, effort and expense needed to increase accuracy by measuring highly redundant data. With respect to resolution, it has been shown that *SnB* substructure phasing works well using the data normally measured in a MAD experiment (*e.g.* 2.8–3.0 Å) and that data sets truncated even to 5 Å can lead to solutions. Excellent results can be obtained from consideration of single-wavelength peak anomalous difference data alone. This is probably because this approach avoids the introduction of any errors resulting from incorrect scaling between two data sets. However, independent *SnB* analysis of data measured at different wavelengths can provide valuable confirmation for questionable sites.

The analyses of data processing (*DREAR*) and *SnB* parameters reported here suggest a set of default values for determining Se-atom substructures; these defaults have been incorporated into *SnB* v2.0. Thus, substructure phasing can be performed almost automatically or the user can intervene if desired. The program can be downloaded from http://www.hwi.buffalo.edu/SnB.

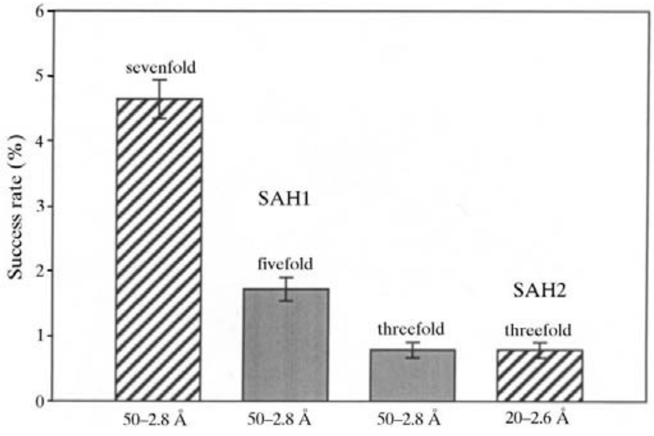


Figure 12
Plot of *SnB* success rates for SAH1 and SAH2 for the standard *SnB* test (black strips) and after truncating the SAH1 data to be either threefold or fivefold redundant (solid gray). Error bars are drawn at ±1σ.

This work is supported by grants from the Medical Research Council of Canada to PLH (MT 13337), from the National Institutes of Health to PLH (GM 29332) and to GDS, RBH and CMW (GM46733) and from the National Science Foundation to CMW (ACI-9721373). The authors thank R. T. Borchardt, C.-S. Yuan and X. Yang for help with the production of the selenomethionyl protein and M. A. Turner for the production of the crystals, data collection and data processing. The X-ray diffraction data were collected at Station X12-C, Brookhaven National Laboratory, a DOE-supported facility.

References

Anderson, D. S., Weiss, M. S. & Eisenberg, D. (1996). *Acta Cryst.* D**52**, 469–480.

Athappilly, F. K. & Hendrickson, W. A. (1995). *Structure*, **3**(12), 1407–1419.

Blessing, R. H. (1989). *J. Appl. Cryst.* **22**, 396–397.

Blessing, R. H. (1997*a*). *J. Appl. Cryst.* **30**, 176–178.

Blessing, R. H. (1997*b*). *J. Appl. Cryst.* **30**, 421–426.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Acta Cryst.* D**52**, 257–266.

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

Brunger, A. T., Adams, P. D., Clore, G. M., Gros, P., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–931.

Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. A. (1997). *Acta Cryst.* A**53**, 436–444.

Cochran, W. (1995). *Acta Cryst.* **8**, 473–478.

Deacon, A. M., Ni, Y., Coleman, W. G. Jr & Ealick, S. E. (1999). Am. Cryst. Assoc. Conference, PT21.

Deacon, A. M., Weeks, C. M., Miller, R. & Ealick, S. E. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.

Doublie, S., Tabor, S., Long, A. M., Richardson, C. C. & Ellenberger, T. (1998). *Nature (London)*, **391**, 251–258.

French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.

Germain, G., Main, P. & Woolfson, M. M. (1971). *Acta Cryst.* A**27**, 368–376.

Hendrickson, W. A. (1985). *Trans. Am. Cryst. Assoc.* **21**, 11–18.

Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.

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

Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Cryst.* A**31**, 480–497.

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

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

Nagar, B., Jones, R. G., Diefenbach, R. J., Isenman, D. E. & Rini, J. M. (1998). *Science*, **280**(5367), 1277–1281.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Prive, G. G., Anderson, D. H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). *Protein Sci.* **8**(7), 1400–1409.

Ramakrishnan, V. & Biou, V. (1997). *Methods Enzymol.* **276**, 538–557.

Sheldrick, G. M. (1982). *Crystallographic Computing*, edited by D. Sayre, pp. 506–514. Oxford: Clarendon Press.

Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.

Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Acta Cryst.* D**53**, 551–557.

Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. (1998). *Acta Cryst.* D**54**, 799–804.

Terwilliger, T. C., Kim, S.-H. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 1–5.

Thanos, C. D., Goodwill, K. E. & Bowie, J. U. (1999). *Science*, **283**, 833–836.

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

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

Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Acta Cryst.* D**49**, 179–181.

Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.

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

Xu, H., Weeks, C. M., Deacon, A. M., Miller, R. & Hauptman, H. A. (2000). *Acta Cryst.* A**56**, 112–118.

Yang, W., Hendrickson, W. A., Crouch, R. J. & Satow, Y. (1990). *Science*, **249**, 1398–405.