Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Optimizing statistical *Shake-and-Bake* for Se-atom substructure determination

Hongliang Xu, Charles M. Weeks and Herbert A. Hauptman

Hauptman–Woodward Medical Research Institute and Department of Structural Biology, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, 73 High Street, Buffalo, NY 14203, USA

Received 22 December 2004
Accepted 13 April 2005

A novel statistical approach to the phase problem in X-ray crystallography was introduced in a recent paper [Xu & Hauptman (2004), *Acta Cryst.* A**60**, 153–157]. In this approach, a new minimal function based on the statistical distribution of structure-invariant values serves as the foundation of an optimization procedure called statistical *Shake-and-Bake*. Favorable application of this procedure to Se-atom substructure determination depends on the choice of the statistical interval over which the function is defined. The effects of interval variation have been studied for 19 Se-atom substructures ranging in size from five to 70 Se atoms in the asymmetric unit and the results have shown an overall improvement in success rate relative to traditional *Shake-and-Bake*. Statistical *Shake-and-Bake* is being incorporated as the default optimization procedure in newly distributed versions of the *SnB* and *BnP* computer programs.

1. Introduction

The difficulty in recovering phase information from measurements of intensities alone is called the 'phase problem'. So far, successful procedures for performing this have been based on the tangent formula (Karle & Hauptman, 1956), the minimal principle (Debaerdemaeker & Woolfson, 1983), maximum entropy (Bricogne, 1984) and minimum charge (Elser, 1999). In the minimal principle method, the phase problem is formulated in terms of constrained global minimization. The problem is complex because of the existence of multiple local minima in the underlying optimization formulations and the success of the minimal principle method depends largely on the radius of convergence of the minimal function. In the last few years, different types of minimal functions, such as the exponential type (Hauptman *et al.*, 1999) and the sine-enhanced type (Xu *et al.*, 2002), have been proposed.

Shake-and-Bake (Weeks *et al.*, 1994) is a multi-solution or multi-trial direct-methods procedure that automatically and repetitively alternates reciprocal-space phase refinement (Shaking) with a complementary real-space density modification to impose physical constraints (Baking). The phase-refinement portion of the *Shake-and-Bake* cycle utilizes either the tangent formula or the technique of parameter shift (Bhuiya & Stanley, 1963) to reduce the value of a minimal function. *Shake-and-Bake* is a powerful procedure 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. It has also yielded solutions for heavy-atom protein substructures containing as many as 160

doi:10.1107/S0907444905011558

Se atoms (von Delft & Blundell, 2002) 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), *BnP* (Weeks *et al.*, 2002) and *SHELXD* (Sheldrick, 1997, 1998).

1.1. Cosine minimal function

If **H** is an arbitrary reciprocal-lattice vector, then the phase φ_H of the normalized structure factor E_H is defined by

$$E_H = |E_H| \exp(i\varphi_H). \qquad (1)$$

For every pair of reciprocal-lattice vectors (**H**, **K**), the structure invariant (triplet) φ_{HK} is defined by means of

$$\varphi_{HK} = \varphi_H + \varphi_K + \varphi_{-H-K}. \qquad (2)$$

In the traditional probabilistic approach, the atomic position vectors r of the atoms in a crystal are assumed to be random variables uniformly and independently distributed in the unit cell. Standard methods of mathematical probability are applied to derive conditional probability distributions of the structure invariants assuming that the magnitudes $|E|$ are known (Cochran, 1955). Traditional *Shake-and-Bake* implements the cosine minimal function (DeTitta *et al.*, 1994),

$$R(\varphi) = \left(\sum_{H,K} A_{HK}\right)^{-1} \sum_{H,K} A_{HK}\left[\cos(\varphi_{HK}) - \frac{I_1(A_{HK})}{I_0(A_{HK})}\right]^2, \qquad (3)$$

where $A_{HK} = 2N^{-1/2}|E_H E_K E_{H+K}|$, N is the number of non-H atoms in the unit cell and I_m, $m = 0, 1$, are modified Bessel functions of order 0 and 1. The cosine minimal function measures the mean-square difference between the values of the cosine structure invariants, $\cos(\varphi_{HK})$, for a set of trial phases and their conditional expected values, $I_1(A_{HK})/I_0(A_{HK})$, derived from probability theory.

1.2. Statistical minimal function

In a recent paper (Xu & Hauptman, 2004), a novel statistical approach to the phase problem in X-ray crystallography was introduced. This approach takes full advantage of the statistical properties of the structure invariants to construct a novel statistical maximal/minimal function. Let $I = [-r, r] \subset [-\pi, \pi]$ be an arbitrary statistical interval, N_I the number of triplets whose values are in I and N_T the total number of triplets. The statistical maximal function is then defined by

$$M(\varphi) = N_I/N_T \qquad (4)$$

and the statistical minimal function is defined by means of

$$m(\varphi) = 1 - (N_I/N_T). \qquad (5)$$

Note that N_I is an implicit function of all selected phases. When an individual phase value changes, all triplet values associated with this phase will change and therefore the value of N_I will also change. It is obvious that the values of the statistical minimal function depend on the choice of the statistical interval $I = [-r, r]$. It was anticipated (and later confirmed experimentally) that with a proper choice of the statistical interval the statistical minimal function reaches its constrained global minimum when all phases are equal to their true values for any choice of origin and enantiomorph (statistical minimal principle). The initial applications of statistical *Shake-and-Bake*, a modification of traditional *Shake-and-Bake* obtained by replacing the cosine minimal function (3) by the statistical minimal function (5), have shown that the statistical approach to the phase problem is a simple, reliable, less computationally intensive and more efficient procedure for the determination of both centrosymmetric and non-centrosymmetric structures, including heavy-atom substructures (Xu & Hauptman, 2004). Owing to the successful direct-methods applications that utilize anomalous dispersion measurements or multiple diffraction patterns (SIR, SAS and MAD) to determine heavy-atom substructures, we focus our attention on optimizing statistical *Shake-and-Bake* for substructure determination.

2. Materials and methods

Both traditional *Shake-and-Bake* and statistical *Shake-and-Bake* were applied to 19 known Se-atom substructures ranging in size from five to 70 Se atoms in the asymmetric unit using a modified version of the computer program *SnB* (Weeks & Miller, 1999). Basic facts regarding the structures, such as the PDB code, number of Se atoms in the asymmetric unit, space group and data resolution, are listed in Table 1. Three-wavelength MAD (multi-wavelength anomalous dispersion) data were available for each structure and in each case *SnB* applications designed to locate the positions of the substructure atoms were made using both the peak-wavelength anomalous differences (PK_{ano} data) and the isomorphous dispersive differences between the inflection point and high-energy remote wavelengths (IP_{iso} data). The normalized difference structure-factor magnitudes, $|E_\Delta|$, were calculated with a series of programs from Blessing's data-reduction and error-analysis routines (*DREAR*): *LEVY* and *EVAL* for structure-factor normalization (Blessing *et al.*, 1996), *LOCSCL* for local scaling of the SIR and SAS magnitudes (Blessing, 1997) and *DIFFE* for computing the actual SIR and SAS difference magnitudes (Blessing & Smith, 1999).

A sample of 1000 randomly positioned N-atom trial structures (where N is the number of independent Se atoms in the asymmetric unit) was generated for each substructure and the dual-space *SnB* refinement procedure was applied to each. The default values of the important size-dependent *SnB* parameters (including the numbers of phases, triplets, *SnB* cycles and peaks selected) that were used in these experiments are summarized in Table 2. These default values are the results of previous analyses of traditional *Shake-and-Bake* applications to Se-atom substructures (Howell *et al.*, 2000; Xu *et al.*, 2002) and they were adopted for statistical *Shake-and-Bake* as well.

Shake-and-Bake belongs to the class of phasing methods known as 'multi-trial' or 'multi-solution' procedures (Germain & Woolfson, 1968). In this study, the comparison of different *Shake-and-Bake* protocols is based on success rate (*i.e.* the

Table 1
Selenium substructure data sets used in this investigation.

		Selenium sites			Resolution (Å)		
Structure ID	PDB code	Theoretical†	Actual‡	Space group	Original§	Actual¶	Reference
PURE	1qcz	5	4	$I422$	1.50	3.00	Mathews *et al.* (1999)
AK	1bx4	8	7	$P2_12_12$	2.25	3.00	Mathews *et al.* (1998)
MTAP	1cb0	9	8	$P321$	2.20	3.00	Appleby *et al.* (1999)
CBAL	1t5h	10	10	$P3_221$	2.50	3.00	Gulick *et al.* (2004)
GAR	1gso	13	13	$P2_12_12_1$	2.22	3.00	Wang *et al.* (1998)
THID	1jxh	14	14	$P4_12_12$	2.30	3.00	Cheng *et al.* (2002)
OMPDC	1dbt	21	19	$P2_12_12$	2.49	3.00	Appleby *et al.* (2000)
SAMDC	1jen	24	22	$P2_1$	2.25	3.00	Ekstrom *et al.* (1999)
MMEPI	1jc4	28	24	$P2_1$	2.00	3.00	McCarthy *et al.* (2001)
AIRS	1cli	28	28	$P2_12_12_1$	3.00	3.00	Li *et al.* (1999)
ADOHCY	1a7a	32	30	$C222$	2.80	3.00	Turner *et al.* (1998)
E1	1l8a	42	40	$P2_1$	2.60	3.00	Arjunan *et al.* (2002)
MUTS	1e3m	48	45	$P2_12_12_1$	3.00	3.00	Lamers *et al.* (2000)
PHI6	1hi8	50	50	$P3_2$	2.80	3.00	Keitel *et al.* (1997)
HYDAN	1gkp	54	54	$C222_1$	2.50	3.00	Abendroth *et al.* (2002)
AEPT	1m32	66	66	$P2_1$	2.55	3.00	Chen *et al.* (2002)
HMGR	1dq8	68	60	$P2_1$	2.33	3.00	Istvan *et al.* (2000)
TRYP	1e2y	70	60	$P2_1$	3.20	3.20	Alphey *et al.* (2000)
AGME	1eq2	70	70	$P2_1$	2.91	3.00	Deacon *et al.* (2000)

† Potential sites based on the amino-acid sequence. ‡ Number of sites reported in the published protein structure. § Measured data resolution. ¶ Truncated data resolution for substructure determination.

Table 2
Default values of *SnB* experimental parameters.

N is the number of Se atoms in the asymmetric unit.

Parameters	$N \leq 10$	$N > 10$
Phases	300	30N
Triplets	3000	300N
Peaks	N	N
Cycles	20	2N

percentage of trial structures that converge to solution). When performing postmortem studies using data for previously known structures, a trial structure subjected to the *Shake-and-Bake* procedure is counted as a solution if there is a close match between the peak positions produced by *Shake-and-Bake* and the true atomic positions for some choice of origin and enantiomorph. Of course, in actual applications to unknown structures potential solutions are identified on the basis of final minimal function values. All experiments were conducted on a network of SGI R10000 workstations at the Hauptman–Woodward Medical Research Institute.

The *Shake-and-Bake* algorithm utilizes the following parameter-shift procedure to reduce the value of the targeted minimal function. Firstly, the phases are sorted in decreasing order with respect to the values of the associated $|E|$ values and initial values for each phase are calculated based on a trial structure having randomly positioned atoms. Beginning with the phase having the largest $|E|$ value, each phase (φ_H) is refined in turn. The values of the minimal function are evaluated four times using phase values of φ_H, $\varphi_H \pm S^\circ$ and $\varphi_H \pm 180^\circ$, where S is a predetermined phase shift (shift size). The minimum of these four values is then found and the phase φ_H is updated accordingly. When consideration of a particular phase is complete, parameter shift proceeds to the next phase using refined values immediately in the subsequent refinement of other phases. The notation PS(S°, k) is used to denote a parameter-shift procedure using shift size S° and k iterations (passes through the phase set) of phase refinement per *Shake-and-Bake* cycle. Based on extensive tests involving a variety of heavy-atom substructures in various space groups, PS(90°, 3) was found to be optimal for heavy-atom substructure determinations (Xu *et al.*, 2002) and was used throughout this investigation.

3. Results

3.1. Optimizing the statistical interval $I = [-r, r]$

The statistical maximal function is defined as the fraction of triplets whose values lie in the statistical interval I and the value of the statistical minimal function is obtained by subtracting this value from unity. When the statistical interval I changes, so does the minimal function value. The main goal of this study was to determine the optimal statistical interval for statistical *Shake-and-Bake*. In order to examine the effects of statistical interval variation, a series of intervals $I = [-r, r]$ with r = 60, 65, 70, ..., 110° and the corresponding success rates obtained from statistical *SnB* experiments are listed in Table 3 for $\mathrm{PK_{ano}}$ difference data sets and in Table 4 for $\mathrm{IP_{iso}}$ difference data sets. For each row in these tables, the three largest success rates are listed as bold numbers. The following can be observed.

(i) The statistical interval is a crucial parameter, especially for large substructures.

(ii) There is a correlation between the optimal statistical interval and the number of Se sites in the asymmetric unit. The size of the optimal interval increases as the number of Se sites

Table 3
Success rates (%) obtained for statistical *Shake-and-Bake* using $\mathrm{PK_{ano}}$ data sets for 17 Se-atom substructures.

Two data sets (HYDAN and TRYP) yielded no solution.

Structure ID	Se sites	Statistical interval I = [−r, r]										
		r = 60°	65°	70°	75°	80°	85°	90°	95°	100°	105°	110°
PURE	5	**30.1**	**29.1**	**24.2**	19.1	17.7	15.3	13.5	12.5	11.4	10.4	9.2
AK	8	13.1	**20.2**	**18.2**	**16.1**	12.4	12.0	11.3	10.3	9.5	9.4	7.3
MTAP	9	0.0	0.1	1.9	**5.6**	**5.4**	4.0	**4.7**	4.1	2.4	2.6	1.8
CBAL	10	0.0	0.5	1.6	**6.4**	**4.2**	**4.7**	**4.2**	3.7	2.3	2.4	1.6
GAR	13	0.0	0.0	0.0	0.02	**0.10**	**0.12**	**0.10**	**0.16**	0.04	0.06	0.02
THID	14	0.0	0.0	0.0	0.18	**1.20**	**0.84**	**0.82**	0.50	0.70	0.30	0.0
OMPDC	21	0.0	0.0	0.1	**7.1**	**9.7**	**6.9**	5.1	4.8	4.3	3.1	2.1
SAMDC	24	0.0	0.8	8.7	**12.5**	**12.3**	**14.2**	12.0	11.4	10.9	11.1	8.2
MMEPI	28	0.2	10.0	26.6	**28.8**	**37.3**	**30.3**	**28.8**	27.8	26.8	25.2	26.0
AIRS	28	0.0	0.0	0.0	0.4	**4.4**	**4.4**	**3.9**	**3.9**	**4.5**	3.7	2.6
ADOHCY	32	0.0	0.0	0.0	0.1	**7.7**	**5.5**	**4.5**	4.2	3.7	2.7	3.5
E1	42	0.0	0.0	0.0	2.4	**4.1**	**2.9**	**3.0**	2.3	1.9	0.7	0.7
MUTS	48	0.0	0.0	0.0	0.0	**9.9**	**6.6**	**6.8**	5.5	3.1	2.9	2.6
PHI6	50	0.0	0.0	0.0	0.0	10.8	**28.2**	**26.6**	**20.2**	16.8	14.2	13.1
AEPT	66	0.0	0.0	0.0	0.0	0.3	2.5	**4.5**	**4.3**	**2.7**	1.4	1.7
HMGR	68	0.0	0.0	0.0	4.9	**21.5**	**24.2**	**24.2**	21.2	20.1	11.3	12.9
AGME	70	0.0	0.0	0.0	0.0	1.0	1.9	**3.4**	**3.0**	**3.5**	2.0	1.3

Table 4
Success rates (%) obtained for statistical *Shake-and-Bake* using $\mathrm{IP_{iso}}$ data sets for 15 Se-atom substructures.

Four data sets (SAMDC, MMEPI, PHI6 and AGME) yielded no solution.

Structure ID	Se sites	Statistical interval I = [−r, r]										
		r = 60°	65°	70°	75°	80°	85°	90°	95°	100°	105°	110°
PURE	5	**15.5**	**16.2**	**16.3**	13.1	11.6	11.2	11.8	10.8	10.5	9.6	7.6
AK	8	16.5	**27.3**	**26.6**	**23.3**	22.8	22.2	19.9	14.9	13.2	13.1	9.9
MTAP	9	2.1	4.0	**7.6**	**8.9**	**7.6**	6.7	6.5	4.5	4.6	4.4	3.5
CBAL	10	2.0	5.8	**9.4**	**10.2**	**7.3**	6.8	6.3	6.5	6.1	4.1	3.9
GAR	13	7.4	12.5	12.6	**13.5**	**12.8**	12.6	**13.8**	11.0	11.5	8.0	6.2
THID	14	9.5	**22.1**	**20.1**	**15.4**	12.7	12.4	11.5	12.1	11.0	11.0	10.3
OMPDC	21	0.7	8.1	**9.1**	8.3	**8.6**	**8.7**	8.2	6.8	5.4	4.1	3.4
AIRS	28	0.0	0.0	0.0	1.4	**4.5**	**4.0**	**4.4**	3.5	4.5	2.7	2.7
ADOHCY	32	0.0	0.0	0.0	0.10	0.70	1.06	1.52	**1.78**	**1.68**	**1.64**	1.52
E1	42	0.0	0.0	0.0	9.3	**12.6**	**11.8**	**13.7**	11.7	10.1	6.9	3.6
MUTS	48	0.5	1.9	4.6	7.1	7.1	6.2	7.0	**11.2**	**11.2**	**9.6**	7.8
HYDAN	54	0.0	0.0	0.0	0.0	1.6	**1.8**	**2.2**	**1.7**	**1.7**	1.0	1.0
AEPT	66	0.0	0.0	0.0	0.3	18.5	18.7	**28.1**	**23.0**	**20.1**	18.3	14.2
HMGR	68	0.0	0.0	0.0	8.1	8.3	8.7	**10.8**	**10.5**	**10.6**	7.5	5.7
TRYP	70	0.0	0.0	0.8	6.5	6.8	8.8	**10.2**	**11.3**	**10.7**	8.8	9.2

increases. This result may be correlated with a sharp distribution of the structure invariants for small substructures and a flat distribution of the structure invariants for large substructures.

(iii) The pattern of bold numbers in Table 3 (for $\mathrm{PK_{ano}}$) is similar to that in Table 4 (for $\mathrm{IP_{iso}}$). Thus, both types of difference data have similar optimal statistical intervals.

(iv) To solve medium or large substructures (≥30 Se atoms), the statistical *Shake-and-Bake* procedure requires a large statistical interval $I = [-r, r]$ with $r \geq 80°$.

(v) Two $\mathrm{PK_{ano}}$ data sets (HYDAN and TRYP) and four $\mathrm{IP_{iso}}$ data sets (SAMDC, MMEPI, PHI6 and AGME) do not yield solutions using any of the statistical intervals that were tested. Nevertheless, all 19 substructures are solvable with a combination of $\mathrm{PK_{ano}}$ and $\mathrm{IP_{iso}}$ data sets.

3.2. Strategy for choosing default statistical interval

It can be observed from Tables 3 and 4 that the statistical interval I = [−90, 90°] yields optimal or near-optimal success rates for medium and large substructures. This interval is not optimal for small substructures; however, the success rates for such substructures are relatively high anyway. Therefore, a conservative strategy is to choose I = [−90, 90°] as the default statistical interval. Based on the results, an aggressive strategy would be to consider the length of the statistical interval, r, as a function of the number of Se atoms, N, in the asymmetric unit of the crystal. From the pattern of bold numbers in Tables 3 and 4, one could assume that $r = a\ln(N) + b$, where a and b are parameters to be determined. After applying least squares to the data (N, r) that produced the bold success rates in Tables 3 and 4, the relationship

$$r = 9.14\ln(N) + 55.31° \qquad (6)$$

is obtained. The empirical formula (6) can be used to calculate the aggressive statistical interval for any targeted substructure provided that the size of the substructure is known.

3.3. Comparison of traditional and statistical *Shake-and-Bake*

The success rates obtained using traditional *Shake-and-Bake* as well as statistical *Shake-and-Bake* with either conservative or aggressive statistical intervals are listed in Table 5 under the headings COS, STAT(C) and STAT(A), respectively, for the $\mathrm{PK_{ano}}$ and $\mathrm{IP_{iso}}$ difference data sets of the 19 Se-atom substructures. Success rates are reported in the form of $x \pm \sigma(x)$, where $\sigma(x)$ is the standard deviation calculated by Bernoulli's distribution, $\sigma(x) = [nx(1 - x)]^{1/2}$, with n being the number of trials and x being the success rate expressed as a fraction. When comparing two success rates (x and y) obtained from two different procedures, y is statistically higher than x if $y \geq x + 2\sigma(|y - x|)$, where $\sigma(|y - x|) = [\sigma^2(x) + \sigma^2(y)]^{1/2}$; y is statistically lower than x if $y \leq x - 2\sigma(|y - x|)$; otherwise y is statistically equivalent to x. When compared with traditional *Shake-and-Bake* using 38 difference data sets, STAT(C) yielded 11 statistically higher and one statistically lower success rates, while STAT(A) yielded 19 statistically higher and one statistically lower success rates.

These results clearly show that statistical *Shake-and-Bake* with either an aggressive or a conservative statistical interval outperforms traditional *Shake-and-Bake* for Se-atom

Table 5
Comparison of success rates (%) obtained from traditional and statistical *Shake-and-Bake* using PK_{ano} and IP_{iso} data sets for 19 Se-atom substructures.

COS represents traditional *Shake-and-Bake*, whereas STAT(C) and STAT(A) represent statistical *Shake-and-Bake* using conservative or aggressive statistical intervals, respectively.

	PK_{ano}			IP_{iso}		
Structure ID	COS	STAT(C)	STAT(A)	COS	STAT(C)	STAT(A)
PURE	11.4 ± 1.01	13.5 ± 1.08	29.8 ± 1.45†‡	10.0 ± 0.95	11.8 ± 1.02	15.7 ± 1.15†
AK	10.2 ± 0.95	11.3 ± 1.00	17.0 ± 1.19†‡	9.0 ± 0.90	19.9 ± 1.26§	26.1 ± 1.39†‡
MTAP	3.4 ± 0.57	4.7 ± 0.67	4.3 ± 0.64	6.1 ± 0.76	6.5 ± 0.78	9.5 ± 0.93†‡
CBAL	3.2 ± 0.56	4.2 ± 0.63	4.1 ± 0.63	6.3 ± 0.77	6.3 ± 0.77	9.1 ± 0.91†‡
GAR	0.0	0.1 ± 0.10	0.0	10.6 ± 0.97	13.9 ± 1.09	11.9 ± 1.02
THID	0.3 ± 0.17	1.0 ± 0.31‡	0.1 ± 0.10	11.6 ± 1.01	11.5 ± 1.01	14.8 ± 1.12†‡
OMPDC	5.0 ± 0.69	5.1 ± 0.70	9.2 ± 0.91†‡	6.8 ± 0.80	8.2 ± 0.87	8.4 ± 0.88
SAMDC	10.4 ± 0.97	12.0 ± 1.03	12.0 ± 1.03	0.0	0.0	0.0
MMEPI	24.6 ± 1.36	28.8 ± 1.43§	31.3 ± 1.47†	0.3 ± 0.17	0.0	0.0
AIRS	2.4 ± 0.48	3.9 ± 0.61	4.9 ± 0.68†	0.8 ± 0.28	4.4 ± 0.65§	4.6 ± 0.66†
ADOHCY	3.3 ± 0.56	4.5 ± 0.65	6.1 ± 0.76†	2.0 ± 0.44	1.6 ± 0.40	1.1 ± 0.33
E1	1.2 ± 0.34	3.0 ± 0.54§	3.4 ± 0.57†	9.9 ± 0.94	13.7 ± 1.09§	13.5 ± 1.08†
MUTS	2.9 ± 0.53	6.8 ± 0.80§	5.0 ± 0.69†	11.1 ± 0.99†§	7.0 ± 0.81	7.7 ± 0.84
PHI6	13.3 ± 1.07	26.6 ± 1.40§	27.1 ± 1.41†	0.0	0.0	0.0
HYDAN	0.0	0.0	0.0	1.1 ± 0.33	2.2 ± 0.46	2.1 ± 0.45
AEPT	2.6 ± 0.50	4.5 ± 0.66‡§	2.7 ± 0.51	15.8 ± 1.15	28.1 ± 1.42‡§	22.4 ± 1.32†
HMGR	14.6 ± 1.12	24.2 ± 1.35§	23.0 ± 1.33†	7.0 ± 0.81	10.8 ± 0.98§	12.6 ± 1.05†
TRYP	0.0	0.0	0.0	13.0 ± 1.06	10.2 ± 0.96	10.6 ± 0.97
AGME	2.1 ± 0.45	3.4 ± 0.57	2.9 ± 0.53	0.0	0.0	0.0

† There is a statistically significant difference in success rates between COS and STAT(A). ‡ There is a statistically significant difference in success rates between STAT(A) and STAT(C). § There is a statistically significant difference in success rates between COS and STAT(C).

substructure determination. The advantage of using a conservative statistical interval is that one fixed interval can be used to determine substructures of any size with a reasonably high success rates. The disadvantage is the loss of higher success rates for small and medium substructures (5–35 Se atoms). The data in Tables 3 and 4 suggest that a large statistical interval, $I = [-r, r]$ with $r > 100°$, may be needed to determine very large substructures (≥100 atoms). On the other hand, the advantage of using an aggressive statistical interval is the potential of yielding maximal success rates for small substructures.

3.4. Effects of measurement errors

As shown in Table 5, large differences between the success rates for PK_{ano} and IP_{iso} difference data were observed for seven test substructures (GAR, SAMDC, MMEPI, PHI6, HYDAN, TRYP and AGME). To investigate the possible causes of these differences, the effects of data accuracy were studied for PK_{ano} from GAR and IP_{iso} from MMEPI by applying statistical *Shake-and-Bake* to error-free data generated using the program *EGEN* (R. Blessing, personal communication) and the known Se atomic coordinates. In both cases, the results (Table 6) show that the success rates are much higher for the error-free data than the corresponding experimental data, thereby indicating that experimental error is in fact the cause of the low success rates. It should be noted that the same number of reflections (30*N*, where *N* is the number of independent Se atoms) were involved as were used in the corresponding application to real data. However, the identities (Miller indices) of the reflections were different because the *Shake-and-Bake* calculations were carried out with either the largest experimental or error-free normalized difference magnitudes.

Table 6
Comparison of success rates (%) for experimental and error-free data.

Data	GAR, PK_{ano}	MMEPI, IP_{iso}
Experimental	0.1	0.0
Error-free	11.8	23.7

4. Conclusion and discussion

The results described above confirm that statistical *Shake-and-Bake* is more powerful than traditional *Shake-and-Bake* for the determination of Se-atom substructures. Consequently, the statistical *Shake-and-Bake* procedure has been implemented as the default method in the latest versions of the computer programs *SnB* and *BnP*. These programs can be downloaded from the websites http://www.hwi.buffalo.edu/SnB/ and http://www.hwi.buffalo.edu/BnP/, respectively.

The statistical *Shake-and-Bake* adopts the default *SnB* parameters for routine applications. However, some changes in the default values of *SnB* parameters, such as the numbers of reflections and invariants, the minimal $|E|/\sigma(|E|)$ and the number of cycles, may help in solving non-routine and difficult structures (Wang & Ealick, 2003). The changes in the default *SnB* parameters may influence the choice of statistical interval.

In the minimal principle method, the phase problem is formulated as a problem in constrained global minimization.

The 'minimal principle', which asserts that a certain objective function is minimized only by the crystal structure, is employed to solve the phase problem. The probabilistic formulation of the cosine minimal function (3) is a nonconvex nonlinear optimization problem. In the special case of centrosymmetric structures, the probabilistic minimal function can be reformulated into an integer linear programming problem (Vaia & Sahinidis, 2003). This formulation is solvable by well established combinatorial optimization techniques that are guaranteed to provide the global optimum in a finite number of steps without explicit enumeration of all possible combinations of phases. This approach yields a fast and reliable method that resolves the crystallographic phase problem for the case of centrosymmetric structures. However, the probabilistic minimal function can not be reformulated as an integer linear programming problem for non-centrosymmetric structures. With the introduction of the statistical approach, the phase problem of non-centrosymmetric structures can now be reformulated into an integer linear programming problem, and experiments designed to test this formulation are now under way.

This research was supported by NIH grants EB002057 and GM-72023. We thank people for making their MAD data available to us and referees for useful suggestions.

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.

Bhuiya, A. K. & Stanley, E. (1963). *Acta Cryst.* **16**, 981–984.

Blessing, R. H. (1997). *J. Appl. Cryst.* **30**, 176–177.

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.

Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.

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.

Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.

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 & Blundell, T. L. (2002). *Acta Cryst.* A**58**, C239.

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.

Elser, V. (1999). *Acta Cryst.* A**55**, 489–499.

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.

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

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

Howell, P. L., Blessing, R. H., Smith, G. D. & Weeks, C. M. (2000). *Acta Cryst.* D**56**, 604–617.

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

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

Keitel, T., Kramer, A., Wessner, H., Scholz, C., Schneider-Mergener, J. & Hohne, W. (1997). *Cell*, **91**, 811–820.

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.

Sheldrick, G. M. (1997). *Prooceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. S. Ashton & S. Bailey, pp. 147–158. Warrington: Daresbury Laboratory.

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

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.

Vaia, A. & Sahinidis, N. V. (2003). *Acta Cryst.* A**59**, 452–458.

Wang, J. & Ealick, S. E. (2003). *J. Appl. Cryst.* **36**, 1397–1401.

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

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.

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.