*Acta Cryst.* (1999). A**55**, 891–900

# *Exponential Shake-and-Bake*: theoretical basis and applications

Herbert A. Hauptman, Hongliang Xu,* Charles M. Weeks and Russ Miller

*Hauptman–Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203, USA. E-mail: xu@hwi.buffalo.edu*

(*Received 16 October 1998; accepted 9 March 1999*)

## Abstract

The simple cosine function used in the formulation of the traditional minimal principle and the related *Shake-and-Bake* algorithm is here replaced by a function of exponential type and its expected value and variance are derived. These lead to the corresponding exponential minimal principle and its associated *Exponential Shake-and-Bake* algorithm. Recent applications of the exponential function to several protein structures within the *Shake-and-Bake* framework suggest that this function leads, in general, to significant improvements in the success rate (percentage of trial structures yielding solution) of the *Shake-and-Bake* procedure. However, only in space group *P*1 is it presently possible to assign optimal values *a priori* for the exponential-function parameters.

## 1. Introduction

The minimal principle, which formulates the phase problem as one of constrained global minimization, was first clearly formulated in 1994 (DeTitta *et al.*, 1994). This principle is the theoretical basis for the *Shake-and-Bake* algorithm (Weeks, DeTitta *et al.*, 1994), which, by alternating phase refinement in reciprocal space with a peak-picking protocol in real space (thus imposing the constraints), has greatly strengthened the traditional techniques of direct methods. During the phase-refinement step, *Shake-and-Bake* typically employs a parameter-shift optimization strategy (Bhuiya & Stanley, 1963) to reduce the value of the cosine minimal function (Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.*, 1994). A variant of this paradigm, alternating tangent refinement (Karle & Hauptman, 1956) in reciprocal space with a peak list optimization technique in real space, and termed *half-baked*, or *SHELX-D*, has been proposed by Sheldrick & Gould (1995). These advances have rendered routine the solution of the phase problem for structures containing as many as 1000 independent non-H atoms in the unit cell, provided that diffraction data to a resolution of 1.2 Å, at least, are available. Their ultimate potential is still unknown.

Our major goal here is to replace the original minimal function, based on the simple cosine, by one of exponential type, in an effort to improve the performance of *Shake-and-Bake*. The traditional minimal principle and the related *Shake-and-Bake* algorithm are replaced by the exponential minimal principle and its associated *Exponential Shake-and-Bake* algorithm. Recent applications of the exponential minimal function to several protein structures within the *Shake-and-Bake* framework show that *Exponential Shake-and-Bake* has the potential, in general, to reduce the time to solution, when compared with traditional *Shake-and-Bake*, by a factor of two or three. The time to solution, however, depends not only on the nature of the minimal function itself but also on the value of the shift angle used in the phase-refinement half of the *Shake-and-Bake* cycle. Therefore, in order to realize the full potential of *Exponential Shake-and-Bake*, it is necessary to determine, prior to structure solution, the best value for the shift angle. A procedure for calculating the optimal value of this shift has been devised only for the space group *P*1. Thus the improvement that *Exponential Shake-and-Bake* promises has so far been realized only for this space group.

## 2. The probabilistic background

By its heavy dependence on the previous work of DeTitta *et al.* (1994), the present analysis is greatly abbreviated. If $\mathbf{H}$ is an arbitrary reciprocal-lattice vector, then the normalized structure factor $E_H$ is defined by

$$E_H = |E_H| \exp(i\varphi_H) = N^{-1/2} \sum_{j=1}^{N} \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \qquad (1)$$

where $N$ is the number of atoms, here assumed for simplicity to be identical, in the unit cell, and $\mathbf{r}_j$ is the position vector of the atom labeled $j$. For every pair of reciprocal-lattice vectors $(\mathbf{H}, \mathbf{K})$, the structure invariant (triplet) $\varphi_{HK}$ is defined by

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

It is assumed that the atomic position vectors $\mathbf{r}_j$ are random variables which are uniformly and independently distributed in the asymmetric unit. Then, for fixed reciprocal-lattice vectors $(\mathbf{H}, \mathbf{K})$, the triplet $\varphi_{HK}$ [equation (2)], as a function [equation (1)] of the primitive random variables $\mathbf{r}_j$, is itself a random variable. The conditional probability distribution, $P(\Phi|A_{HK})$, of the triplet $\varphi_{HK}$, given the three magnitudes

$$|E_H|, \quad |E_K|, \quad |E_{H+K}|, \tag{3}$$

is known to be

$$P(\Phi|A_{HK}) = [1/2\pi I_0(A_{HK})]\exp(A_{HK}\cos\Phi), \tag{4}$$

where $\Phi$ represents the triplet $\varphi_{HK}$,

$$A_{HK} = (2/N^{1/2})|E_H E_K E_{H+K}| \tag{5}$$

and $I_0$ is the modified Bessel function (Cochran, 1955).

## 3. Traditional *Shake-and-Bake*

### 3.1. *The expected value and variance*

From the distribution (4), the conditional expected value and the conditional variance of $\cos(\varphi_{HK})$, given $A_{HK}$, are readily found (*e.g.* DeTitta *et al.*, 1994):

$$\varepsilon[\cos(\varphi_{HK})|A_{HK}] = \frac{I_1(A_{HK})}{I_0(A_{HK})} \tag{6}$$

and

$$\mathrm{var}[\cos(\varphi_{HK})|A_{HK}] = \tfrac{1}{2} + \frac{I_2(A_{HK})}{2I_0(A_{HK})} - \frac{I_1^2(A_{HK})}{I_0^2(A_{HK})}, \tag{7}$$

where $I_0$, $I_1$ and $I_2$ are modified Bessel functions.

### 3.2. *The minimal function and the minimal principle*

Since, as is readily confirmed, $A_{HK}$ is strongly correlated with the reciprocal of the variance [equation (7)], one defines the minimal function, $R(\varphi)$, a function of the phases in view of equation (2), by means of

$$R(\varphi) = \Bigg(\sum_{H,K} A_{HK}\Bigg)^{-1} \sum_{H,K} A_{HK}\left[\cos\varphi_{HK} - \frac{I_1(A_{HK})}{I_0(A_{HK})}\right]^2 \tag{8}$$

and conjectures, in view of equations (6) and (7), that the constrained global minimum of $R(\varphi)$ yields the values of the individual phases for some choice of origin and enantiomorph (the minimal principle). Owing to the existence of identities among the individual phases, we seek the 'constrained' global minimum of $R(\varphi)$, not the unconstrained global minimum; this is an essential distinction that *Shake-and-Bake* exploits.

Although the inverse of the variance, $\mathrm{var}^{-1}$ [equation (7)], could be used instead of $A$ as the weight in equation (8), this substitution does not improve the performance of *Shake-and-Bake*. For this reason, the $A$ values are used as weights in equation (8) rather than the more complicated $\mathrm{var}^{-1}$ values.

### 3.3. *The constrained global minimum, $R_T$, of $R(\varphi)$*

In view of equation (4), it is readily confirmed (*e.g.* DeTitta *et al.*, 1994) that when the phases are set equal to their true values, then, no matter what the choice of origin or enantiomorph, the value $R_T$ of $R(\varphi)$ becomes

$$R_T = \tfrac{1}{2} + \Bigg(\sum_{H,K} A_{HK}\Bigg)^{-1} \sum_{H,K} A_{HK}\left[\frac{I_2(A_{HK})}{2I_0(A_{HK})} - \frac{I_1^2(A_{HK})}{I_0^2(A_{HK})}\right] < \tfrac{1}{2}, \tag{9}$$

which clearly is the constrained global minimum of $R(\varphi)$.

### 3.4. *The value, $R_R$, of $R(\varphi)$ when the phases are chosen at random*

In this case, it is easily seen, as reference to equation (8) shows, that

$$R_R = \tfrac{1}{2} + \Bigg(\sum_{H,K} A_{HK}\Bigg)^{-1} \sum_{H,K} A_{HK}\frac{I_1^2(A_{HK})}{I_0^2(A_{HK})} > \tfrac{1}{2}. \tag{10}$$

### 3.5. *The discriminant D*

From equations (9) and (10), we conclude that

$$R_T < \tfrac{1}{2} < R_R. \tag{11}$$

Equation (5) shows that, as $N$ increases indefinitely, $A$ values tend to become very small, so that, in view of equation (9), $R_T$ approaches $\frac{1}{2}$ from below. Similarly, with increasing $N$, $R_R$ approaches $\frac{1}{2}$ from above. It follows that, as $N$ becomes very large, the value of the discriminant $D$, defined by

$$D = R_T/R_R < 1, \tag{12}$$

approaches unity. In this case, since the constrained global minimum $R_T$ of $R(\varphi)$ approaches $R_R$, one anticipates that structural solution will become more difficult and computer intensive. Of course this is simply another way of saying that more complex structures are harder to solve than less complex ones and justifies regarding $D$ as a measure of the difficulty of structural solution: as $D$ approaches unity, structural solution becomes more difficult and time consuming.

## 4. *Exponential Shake-and-Bake*

We proceed as in the previous section but replace the simple cosine by the exponential

$$g(\varphi_{HK}) = \exp[\mu A_{HK}\cos\eta\cos(\varphi_{HK} + \eta)], \tag{13}$$

dependent on the parameters $\mu$ and $\eta$ (as well as on $A_{HK}$).

### 4.1. *The expected value $\overline{g_{HK}}$ of $g(\varphi_{HK})$*

Referring to equation (4), we find the conditional expected value of $g(\varphi_{HK})$ given $A_{HK}$:

$$\begin{aligned}\overline{g_{HK}} &= \varepsilon[g(\varphi_{HK})|A_{HK}] \\ &= [1/2\pi I_0(A_{HK})]\int\limits_0^{2\pi} \exp[\mu A_{HK}\cos\eta\cos(\Phi+\eta) \\ &\quad + A_{HK}\cos\Phi]\,\mathrm{d}\Phi \qquad (14) \\ &= [1/2\pi I_0(A_{HK})]\int\limits_0^{2\pi} \exp[A_{HK}X\cos(\Phi+\xi)]\,\mathrm{d}\Phi, \qquad (15)\end{aligned}$$

where

$$\begin{aligned}X &= [\mu^2\cos^2\eta + 2\mu\cos^2\eta + 1]^{1/2} \\ &= [\mu(\mu+2)\cos^2\eta + 1]^{1/2} \qquad (16)\end{aligned}$$

and $\xi$ is independent of $\Phi$. Then the integration of (15) is immediate:

$$\overline{g_{HK}} = [1/I_0(A_{HK})]I_0\{A_{HK}[\mu(\mu+2)\cos^2\eta + 1]^{1/2}\}. \qquad (17)$$

### 4.2. *The variance of $g(\varphi_{HK})$*

Replacing $\mu$ by $2\mu$ in equation (17), we find the conditional expected value $\overline{g^2_{HK}}$ of the square of $g(\varphi_{HK})$ given $A_{HK}$:

$$\begin{aligned}\overline{g^2_{HK}} &= \varepsilon[g^2(\varphi_{HK})|A_{HK}] \\ &= [1/I_0(A_{HK})]I_0\{A_{HK}[4\mu(\mu+1)\cos^2\eta + 1]^{1/2}\}. \qquad (18)\end{aligned}$$

Hence the variance of $g(\varphi_{HK})$ is given by

$$\operatorname{var}[g(\varphi_{HK})|A_{HK}] = \overline{g^2_{HK}} - \overline{g_{HK}}^2. \qquad (19)$$

The 'weight' $W_{HK}$ is defined to be the reciprocal of the variance:

$$W_{HK} = \{\operatorname{var}[g(\varphi_{HK})|A_{HK}]\}^{-1}. \qquad (20)$$

### 4.3. *The exponential minimal principle*

In complete analogy to the traditional minimal principle, one now defines the exponential minimal function $m(\varphi)$ by means of

$$m(\varphi) = \left(\sum_{H,K} W_{HK}\right)^{-1} \sum_{H,K} W_{HK}[g(\varphi_{HK}) - \overline{g_{HK}}]^2, \qquad (21)$$

where $\overline{g_{HK}}$ and $W_{HK}$ are defined by equations (17) and (20), respectively, and conjectures that the constrained global minimum of $m(\varphi)$ yields the true values of the phases for some choice of origin and enantiomorph (the exponential minimal principle).

### 4.4. *The constrained global minimum $m_T$ of $m(\varphi)$*

Again, in exact analogy to the derivation of equation (9), we now find the value $m_T$ of $m(\varphi)$ when all phases are equal to their true values for any choice of origin and enantiomorph:

$$m_T = \left(\sum_{H,K} W_{HK}\right)^{-1} \sum_{H,K} W_{HK}(\overline{g^2_{HK}} - \overline{g_{HK}}^2), \qquad (22)$$

which, in view of equations (19) and (20), becomes simply

$$m_T = \left(\sum_{H,K} W_{HK}\right)^{-1} \sum_{H,K} 1, \qquad (23)$$

which is clearly the constrained global minimum of $m(\varphi)$.

### 4.5. *The value $m_R$ of $m(\varphi)$ when the phases are chosen at random*

As in the derivation of equation (10), we now find

$$\begin{aligned}m_R = \left(\sum_{H,K} W_{HK}\right)^{-1} \sum_{H,K} W_{HK}[\overline{g_{HK}}^2 \\ - 2I_0(\mu A_{HK}\cos\eta)\overline{g_{HK}} + I_0(2\mu A_{HK}\cos\eta)]. \qquad (24)\end{aligned}$$

### 4.6. *The exponential discriminant $\Delta$*

In analogy to equation (12), we now define the exponential discriminant $\Delta$ by means of

$$\Delta = m_T/m_R \qquad (25)$$

and infer, from the definitions of $m_T$ and $m_R$, that

$$0 < \Delta < 1. \qquad (26)$$

Furthermore, as with traditional *Shake-and-Bake*, $\Delta$ is a measure of the ease of structural solution: the smaller the value of $\Delta$, the easier it is to solve the structure.

## 5. Materials and methods

Both the cosine minimal function and the exponential minimal function were applied to the series of known structures listed in Table 1 using version 2 of *SnB* (Weeks & Miller, 1999*a*), a computer program that implements *Shake-and-Bake*. Atomic resolution data sets were available for these structures, which range in size from 74 to 1000 non-H protein atoms in the asymmetric unit and crystallize in space groups $P1$, $P2_1$ and $P2_12_12_1$. A sample of 1000 (for small structures) or 500 (for large structures) randomly positioned $n$-atom trial structures (where $n$ is the number of non-H atoms in the asymmetric unit) was generated for each data set. For each structure, an atom:phase:triplet ratio of approxi-

Table 1. *Test data sets used in this investigation*

| Structure | Number of atoms | Space group | Resolution (Å) | Reference |
|---|---|---|---|---|
| Emerimycin | 74 | $P1$ | 0.91 | Marshall *et al.* (1990) |
| Isoleucinomycin | 84 | $P2_12_12_1$ | 0.94 | Pletnev *et al.* (1980) |
| Enkephalin analog | 96 | $P1$ | 0.83 | Krstenansky (unpublished) |
| Ternatin | 104 | $P2_12_12_1$ | 0.94 | Miller *et al.* (1993) |
| Hexaisoleucinomycin | 113 | $P2_12_12_1$ | 1.00 | Pletnev *et al.* (1992) |
| Gramicidin A | 317 | $P2_12_12_1$ | 0.86 | Langs (1988) |
| Crambin | 327 | $P2_1$ | 0.83 | Hendrickson & Teeter (1981) |
| Triclinic vancomycin | 404 | $P1$ | 0.97 | Loll *et al.* (1997) |
| Alpha-1 peptide | 408 | $P1$ | 0.90 | Privé *et al.* (1999) |
| Scorpion toxin II | 508 | $P2_12_12_1$ | 0.96 | Smith *et al.* (1997) |
| Triclinic lysozyme | 1001 | $P1$ | 0.85 | Deacon *et al.* (1998) |

mately 1:10:100 was used in the *Shake-and-Bake* procedure. Values of the basic parameters (*i.e.* the numbers of phases, triplet invariants, peaks and refinement cycles), which are all dependent on structure size (Weeks & Miller, 1999*b*), are summarized in Table 2.

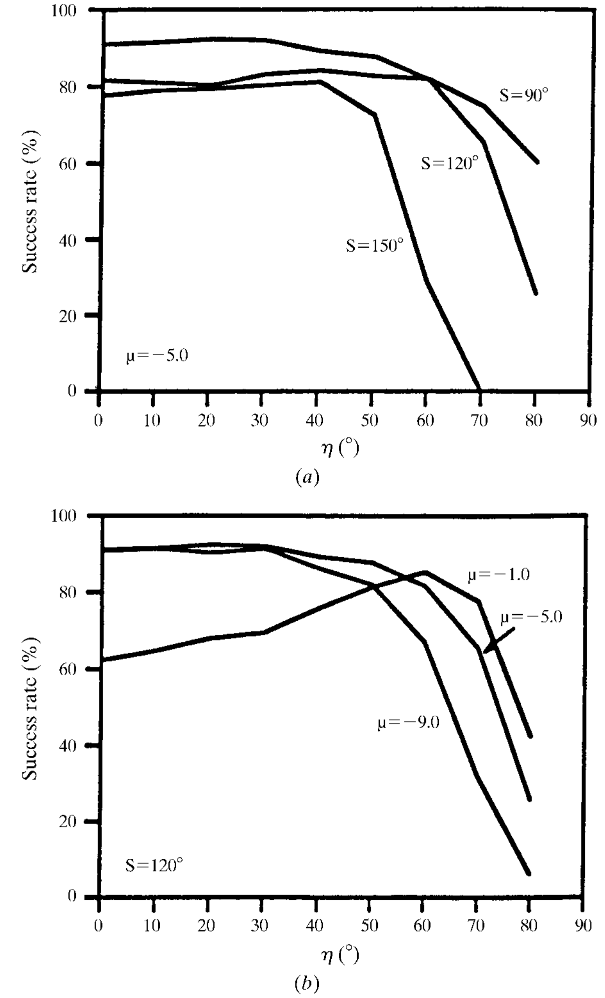


Fig. 1. Success rates of the exponential minimal function as a function of parameter $\eta$ for emerimycin.

The notations $\mathrm{COS}(S, m, k)$ and $\mathrm{EXP}(S, m, k)$ are used to denote parameter-shift optimization of the cosine or exponential minimal functions, respectively, using shift size $S$, a maximum of $m$ steps, and $k$ iterations (passes through the phase set per *Shake-and-Bake* cycle). The notation $\mathrm{PS}(90°, 2)$ denotes the default parameter-shift conditions (*i.e.* the cosine minimal function with a 90° shift size, a maximum of 2 shifts, and 1 iteration for $P1$ structures or 3 iterations for non-$P1$ structures) normally employed in the *SnB* program. These conditions were based on a series of previous studies using data sets for known small-molecule structures (Weeks, Hauptman *et al.*, 1994; Chang *et al.*, 1997). In a recent application of the *Shake-and-Bake* procedure to triclinic lysozyme (Deacon *et al.*, 1998), it was shown that a single large shift of 157.5° produced the best results. Consequently, in this study the cosine minimal function was applied using not only the default conditions [$\mathrm{PS}(90°, 2)$], but also a series of single shifts [$\mathrm{COS}(S, 1, 1)$] with $S = 22.5, 45, 67.5, 90, 112.5, 135, 157.5$ and 180°, in an effort to find the optimal conditions.

In the case of the exponential minimal function [$\mathrm{EXP}(S, 1, 1)$], it is necessary to optimize the parameter-shift angle based on some choice of the parameters $\mu$ and $\eta$. The determination of $\eta$ was based on the information presented in Fig. 1. Fig. 1(*a*) shows success-rate curves for various shift sizes $S$ while $\mu$ was fixed, and Fig. 1(*b*) shows success-rate curves for various values of $\mu$ while shift size $S$ was fixed. It is clear that the optimal value of $\eta$ is between 0 and 40°. Thus, the default value of $\eta$ was set to 20° in an effort to minimize the amount of computation in the investigation.

The exponential minimal function depends on $I_0\{A_{HK}[4\mu(\mu+1)\cos^2\eta+1]^{1/2}\}$, the modified Bessel function. Since the $I_0$ function has an exponential growth rate, one must carefully consider the value of its argument. Assuming that the maximum value of the argument is $X_{\max}$, then

$$A_{HK}[4\mu(\mu+1)\cos^2\eta+1]^{1/2} \le |2\mu+1|A_{\max} \le X_{\max}, \qquad (27)$$

where $A_{\max} = \max_{H,K} A_{HK}$. It follows that

Table 2. *Values of experimental parameters*

| Structure | Phases | Triplets | Peaks | Cycles | Trials |
|---|---|---|---|---|---|
| Emerimycin | 740 | 7400 | 74 | 50 | 1000 |
| Isoleucinomycin | 840 | 8400 | 84 | 100 | 1000 |
| Enkephalin analog | 960 | 9600 | 96 | 100 | 1000 |
| Ternatin | 1040 | 10400 | 84 | 100 | 1000 |
| Hexaisoleucinomycin | 1130 | 11300 | 90 | 125 | 1000 |
| Gramicidin A | 3000 | 30000 | 200 | 300 | 500 |
| Crambin | 3000 | 30000 | 100 | 300 | 500 |
| Triclinic vancomycin | 4000 | 40000 | 150 | 500 | 500 |
| Alpha-1 peptide | 4000 | 40000 | 300 | 500 | 500 |
| Scorpion toxin II | 5000 | 50000 | 200 | 500 | 500 |
| Triclinic lysozyme | 11100 | 111000 | 350 | 750 | 500 |

$$-\tfrac{1}{2}(1+X_{\max}/A_{\max}) \le \mu \le \tfrac{1}{2}(-1+X_{\max}/A_{\max}). \qquad (28)$$

Since $\mu$ cannot equal 0 [otherwise equation (21) degenerates to zero], the range of $\mu$ is

$$-\tfrac{1}{2}(1+X_{\max}/A_{\max}) \le \mu < 0 \qquad (29)$$

and

$$0 < \mu \le \tfrac{1}{2}(-1+X_{\max}/A_{\max}). \qquad (30)$$

On a Silicon Graphics R10000 Indigo workstation, using Fortran77 and storing the result of the computation as a real data type, $X_{\max}$ must be bounded by the value 88.0 in order to avoid overflow. The information presented in Fig. 2 clearly indicates that the optimal value of $\mu$ should be negative. The allowable ranges of $\mu$ for various data sets are shown in Table 3.

Table 3. *The range of $\mu$ from (29) when $X_{max}$ = 88*

| Structure | $A_{\max}$ | Range of $\mu$ |
|---|---|---|
| Emerimycin | 4.70 | (−9.8, 0) |
| Isoleucinomycin | 3.42 | (−13.3, 0) |
| Enkephalin analog | 3.81 | (−12.0, 0) |
| Ternatin | 4.16 | (−11.0, 0) |
| Hexaisoleucinomycin | 4.10 | (−11.2, 0) |
| Gramicidin A | 5.64 | (−8.3, 0) |
| Crambin | 2.26 | (−20.0, 0) |
| Triclinic vancomycin | 2.34 | (−19.3, 0) |
| Alpha-1 peptide | 3.74 | (−12.2, 0) |
| Scorpion toxin II | 1.32 | (−33.8, 0) |
| Triclinic lysozyme | 2.34 | (−19.3, 0) |

In this study, alternative computational procedures are compared on the basis of two criteria. When performing *post mortem* 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 minimal function values. The success rate is defined as the percentage of trial structures that go to solution, and the measurement of success rates at the end of a fixed number of cycles provides one important indication as to the quality of a particular refinement method. However, this measurement by itself provides an incomplete comparison since it does not take into account the computational effort (running time) needed to produce the solutions. The relative efficiency of two methods can be compared as a function of cycle on the basis of the cost effectiveness (CE),

$$\mathrm{CE} = 3600B/TCt, \qquad (31)$$

where $T$ is the number of trial structures, $C$ is the number of cycles per trial structure, $B$ is the number of solutions produced by $T$ such trials, and $t$ is the running time (in s) for one cycle of one trial. In this communication, CE has units of solutions per hour on a Silicon Graphics R10000 Indigo workstation. All experiments were conducted either on a network of SGI R10000 workstations at the Hauptman–Woodward Medical Research Institute, on an IBM SP2 at the Cornell Theory Center (CTC), or on an IBM SP2 at the Center for Computational Science and Technology (CCST) at Argonne National Laboratory.

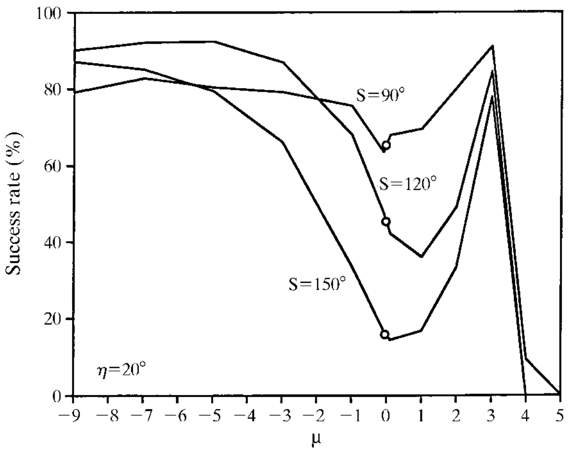


Fig. 2. Success rates of the exponential minimal function as a function of parameter $\mu$ for emerimycin.

## 6. Results

### 6.1. *Traditional cosine minimal function*

Table 4 summaries the *Shake-and-Bake* success rates of the traditional cosine minimal function, COS($S$, 1, 1), for various parameter-shift angles. This table provides a basis for comparing the results of the cosine minimal function with those of the exponential minimal function. It can be observed that when the cosine minimal function is employed with a single shift in the phase-refinement procedure, $S$ = 90° is the optimal (or nearly optimal) parameter-shift size for small structures (emerimycin, isoleucinomycin and enkephalin analog), $S$ = 112.5° is the optimal shift for medium structures (ternatin, hexaisoleucinomycin, gramicidin A, crambin, triclinic vancomycin and alpha-1 peptide) and $S$ = 157.5° is the optimal shift for the largest structure (triclinic lysozyme). In several cases, the success rate for the optimal single shift significantly exceeds that for the default double 90° shift, which appears to be best suited for smaller structures.

### 6.2. *Exponential minimal function*

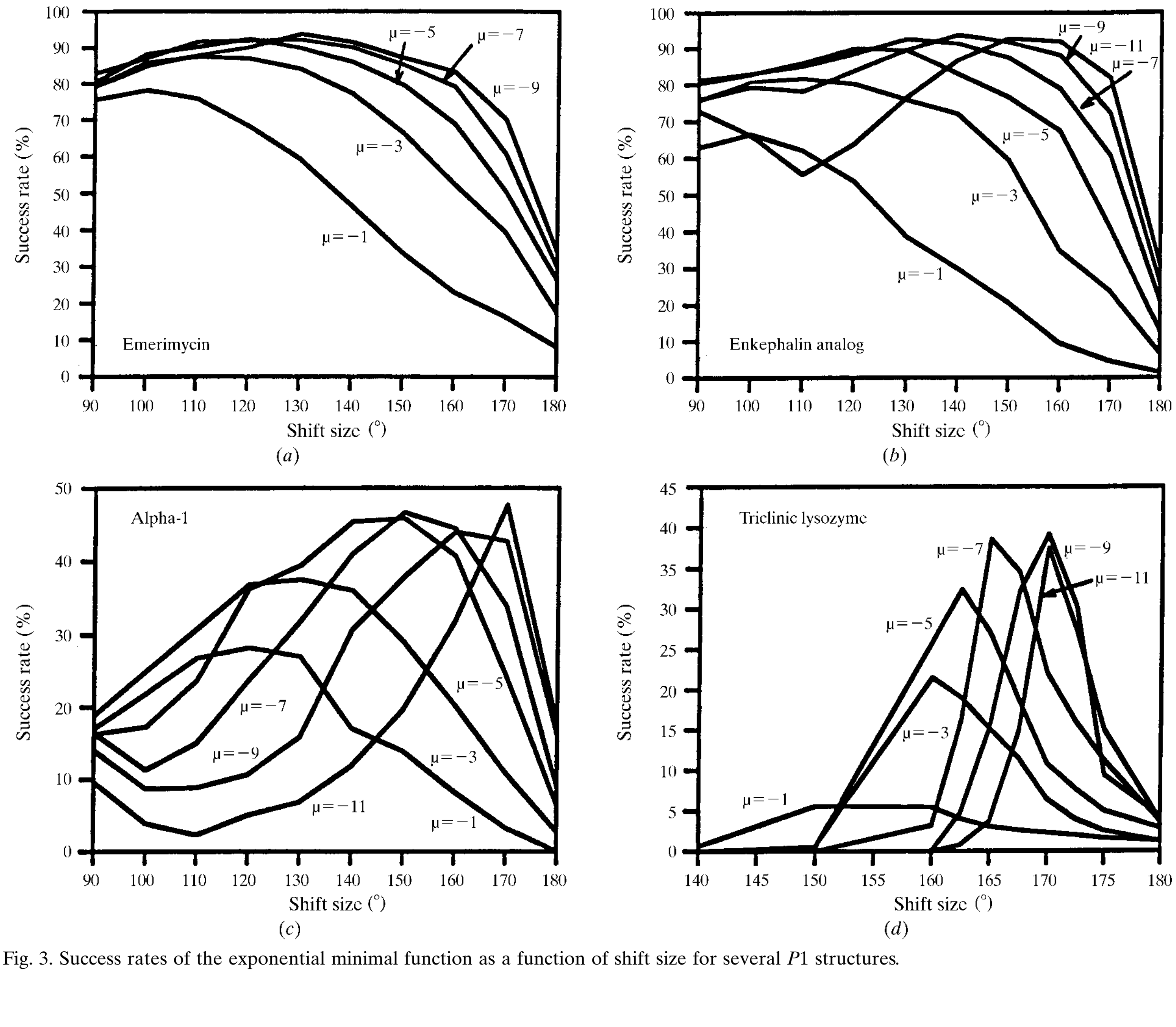
Fig. 3 illustrates success rate as a function of parameter-shift size $S$ when the exponential minimal function is employed in the phase-refinement procedure. The family of curves presented for four structures in space group $P1$ (triclinic vancomycin was omitted because of its very low success rate) shows the results for various values of $\mu$ chosen from Table 3. It can be observed from Fig. 3 that:

(*a*) the optimal parameter-shift size (which leads to the highest success rate) increases when $\mu$ decreases;

(*b*) the sharpness of the success-rate curves depends on the selection of $\mu$ (*i.e.* the more negative the value of $\mu$, the sharper the success-rate curve);

(*c*) the highest success rate of each structure occurs when $\mu$ is near the middle point of the interval given in Table 3;

(*d*) appropriate choices of the parameter-shift size $S$ and the $\mu$ value are critical for achieving an optimal success rate;



(*a*)

(*b*)

(*c*)

(*d*)

Fig. 3. Success rates of the exponential minimal function as a function of shift size for several $P1$ structures.

Table 4. *Success rates (%) of the cosine minimal function*

An asterisk (*) indicates the optimal single-shift size for each structure.

| | Default | Cosine single parameter-shift size COS($S$, 1, 1) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Structure | PS(90°, 2) | 22.5° | 45.0° | 67.5° | 90.0° | 112.5° | 135.0° | 157.5° | 180.0° |
| Emerimycin | 63.1 | 2.9 | 38.1 | 58.0 | 61.6* | 42.2 | 21.7 | 6.7 | 2.2 |
| Isoleucinomycin | 10.4 | 0.0 | 1.2 | 6.2 | 9.9 | 10.2* | 6.0 | 4.8 | 1.9 |
| Enkephalin analog | 43.0 | 0.1 | 14.1 | 35.1 | 35.7* | 24.4 | 8.8 | 2.7 | 0.5 |
| Ternatin | 0.9 | 0.0 | 0.0 | 0.2 | 1.0 | 1.0* | 0.1 | 0.5 | 0.0 |
| Hexaisoleucinomycin | 2.6 | 0.0 | 0.2 | 0.8 | 2.4 | 2.7* | 1.7 | 1.0 | 0.4 |
| Gramicidin A | 1.0 | 0.0 | 0.0 | 0.2 | 1.6 | 2.2* | 1.5 | 0.4 | 0.0 |
| Crambin | 4.8 | 0.0 | 1.0 | 4.4 | 4.2 | 5.2* | 4.4 | 1.0 | 1.4 |
| Triclinic vancomycin | 0.05 | 0.0 | 0.0 | 0.10 | 0.05 | 0.25* | 0.10 | 0.0 | 0.0 |
| Alpha-1 peptide | 13.7 | 0.0 | 0.0 | 2.8 | 15.8 | 19.8* | 8.7 | 3.3 | 0.2 |
| Scorpion toxin II | 1.6 | 0.0 | 0.0 | 0.6 | 1.0 | 0.0 | 1.0* | 0.4 | 0.0 |
| Triclinic lysozyme | 0.0 | 0.0 | 0.6 | 0.2 | 0.0 | 0.0 | 0.8 | 13.5* | 1.6 |

Table 5. *Success rates (%) of the exponential minimal function for the P1 test structures*

An asterisk (*) indicates the optimal parameter-shift size.

| Shift size | Emerimycin | Enkephalin analog | Alpha-1 peptide | Triclinic lysozyme |
|---|---|---|---|---|
| ($S$) (°) | ($\mu$ = −5.0) | ($\mu$ = −6.0) | ($\mu$ = −6.0) | ($\mu$ = −10.0) |
| 90 | 80.4 | 80.4 | 16.2 | 0.0 |
| 100 | 88.3 | 82.8 | 17.2 | 0.0 |
| 110 | 90.0 | 86.0 | 23.6 | 0.0 |
| 120 | 92.4* | 90.1* | 36.2 | 0.0 |
| 130 | 89.8 | 89.5 | 39.4 | 0.0 |
| 140 | 85.9 | 83.0 | 45.4 | 0.0 |
| 150 | 79.4 | 76.7 | 45.8* | 0.0 |
| 160 | 68.7 | 67.3 | 40.6 | 0.0 |
| 170 | 50.3 | 41.0 | 24.2 | 40.2* |
| 180 | 26.4 | 12.9 | 6.4 | 4.2 |

(*e*) replacing the cosine minimal function with the exponential minimal function in the parameter-shift phase-refinement procedure leads, in general, to significant improvements in the success rate of the *Shake-and-Bake* procedure for structures in space group $P1$, provided that appropriate choices of $\mu$, $\eta$ and the parameter-shift size $S$ are made.

Table 5 summarizes the exponential minimal function success rates of four $P1$ structures obtained using various parameter-shift angles and optimal values of parameter $\mu$ near the midpoint of its range for each structure. Based on the information presented in Fig. 3 and Table 5, the following optimal parameters can be suggested for the exponential minimal function:

$$\eta \simeq 20^\circ, \qquad (32)$$

$$\mu \simeq -\tfrac{1}{4}(1 + X_{\max}/A_{\max}), \qquad (33)$$

$$S \simeq 47.8 + 17.3\ln(n), \qquad (34)$$

where $n$ is the number of independent non-H protein atoms. Equation (34) is derived from the least-squares method using the relationship between the optimal parameter-shift size (indicated by an asterisk in Table 5) and the number of independent non-H atoms ($n$) for each $P1$ structure.

### 6.3. *Comparison of methods*

A comparison of success rate and cost effectiveness for the traditional cosine minimal function using the optimal shift size with the results for the exponential minimal function using optimal parameters is presented in Table 6 for five $P1$ structures. It should be pointed out that the experimental data set for triclinic vancomycin (80.2% completeness at 0.97 Å) has been replaced with the error-free calculated data set (100% completeness at 0.97 Å) due to the very low success rate obtained when this data set is incomplete. These data show that using the exponential minimal function with optimal parameters given by equations (32)–(34) leads to significant improvements in both the success rate and cost effectiveness of the *Shake-and-Bake* procedure (with the minor exception of emerimycin, for which the default double 90° shift with the cosine minimal function is most cost effective).

A similar comparison of results for six non-$P1$ structures is presented in Table 7. In this case, the results are inconsistent. It appears that the exponential minimal

Table 6. *Comparison of success rates and cost effectiveness for cosine and exponential minimal functions for P1 structures using optimal parameters for a single shift*

An asterisk (*) indicates the best result.

| | Success rate (%) | | | Cost effectiveness (solutions h$^{-1}$) | | |
|---|---|---|---|---|---|---|
| *P*1 structure | PS(90°, 2) | COS | EXP | PS(90°, 2) | COS | EXP |
| Emerimycin | 63.1 | 61.6 | 92.4* | 618.4* | 494.8 | 600.7 |
| Enkephalin analog | 43.0 | 35.7 | 90.1* | 172.3 | 105.8 | 209.7* |
| Alpha-1 peptide | 13.7 | 19.8 | 45.8* | 0.96 | 0.94 | 2.21* |
| Triclinic lysozyme | 0.0 | 13.5 | 40.2* | 0.0 | 0.14 | 0.39* |
| Triclinic vancomycin (calculated data set) | 1.8 | 14.4 | 41.6* | 0.23 | 0.88 | 2.35* |

Table 7. *Comparison of success rates and cost effectiveness for cosine and exponential minimal functions for $P2_1$ and $P2_12_12_1$ structures*

Default exponential-function parameters were determined using equations (32)–(34). An asterisk (*) indicates the best results for the three computational procedures. The numbers in parentheses indicate the best possible results for the exponential function, not necessarily restricting the parameters to the values given by equations (32)–(34).

| | Success rate (%) | | | Cost effectiveness (solutions h$^{-1}$) | | |
|---|---|---|---|---|---|---|
| *P*1 structure | PS (90°, 2) | COS | EXP | PS (90°, 2) | COS | EXP |
| Isoleucinomycin | 10.4 | 10.2 | 16.3* (24.3) | 20.8 | 18.8 | 33.9* (41.7) |
| Ternatin | 0.9 | 1.0* | 0.0 (1.2) | 1.08 | 1.16* | 0.0 (1.23) |
| Hexaisoleucinomycin | 2.6 | 2.7* | 0.0 (6.0) | 1.44 | 1.45* | 0.0 (3.14) |
| Gramicidin A | 1.0 | 2.2 | 3.2* (6.6) | 0.11 | 0.19 | 0.26* (0.54) |
| Crambin | 4.8 | 5.2 | 5.6* (7.4) | 0.78 | 0.82 | 0.92* (1.10) |
| Scorpion toxin II | 1.6* | 1.0 | 0.0 (2.2) | 0.036* | 0.018 | 0.0 (0.044) |

function parameters given by equations (32)–(34) are not optimal for $P2_1$ and $P2_12_12_1$ structures.

### 6.4. *Implications of the results*

Structures in space group *P*1 exhibit behavior that, in many respects, differs from that of structures crystallizing in other space groups. As shown by the results in Tables 5 and 6, the success rate and cost effectiveness of structures in space group *P*1 are unexpectedly high for both minimal functions. Although no rigorous explanation can be given to explain this observation, it can be argued heuristically that, since it is only in this space group that all origins are permissible, it is most likely that an arbitrarily chosen trial structure will have the correct relative positions for several atoms (for some choice of origin) if the space group is *P*1. This observation then raises the question of whether or not it would be better to treat all structures as if they were *P*1 structures. To answer this, both minimal functions were applied to the 84-atom isoleucinomycin ($P2_12_12_1$) and 327-atom crambin ($P2_1$) structures, treating them as if the space group were *P*1. This required a fourfold increase in computational effort for isoleucinomycin, but only a twofold increase was required for crambin. Success rates and cost effectiveness in *P*1 and in the actual space groups are compared in Table 8. For isoleucinomycin, the best success rate is obtained when the exponential minimal function is applied in space group *P*1; however, the optimum cost effectiveness is obtained when the exponential minimal function is applied in space group $P2_12_12_1$. For crambin, the best success rate and cost effectiveness are both obtained when the exponential minimal function is applied in the space group *P*1.

In their implementation of iterative peak list optimization, a procedure that is closely related to *Shake-and-Bake* and uses additional criteria for peak selection but employs only tangent-formula (Karle & Hauptman, 1956) phase refinement, Sheldrick & Gould (1995) have found it advantageous to treat all structures in space group *P*1. Since these authors also use a substructure model to provide starting coordinates when the structure contains a relatively rigid fragment, working in *P*1 is particularly advantageous since only a rotational search is required. In all other space groups, both rotational and translational searches are necessary.

## 7. Conclusions

In view of the experiments described above, it is recommended that single-shift *Exponential Shake-and-Bake*, with parameters $\eta$, $\mu$ and $S$ defined by equations (32)–(34), be used for structures in space group *P*1. Optimal values of the parameters $\eta$, $\mu$ and $S$ are not yet

Table 8. *Comparison of the success rates and cost effectiveness in P1 and in the actual space group for isoleucinomycin and crambin*

The parameters for the exponential minimal function given by equations (32)–(34) are applied to both space groups. An asterisk (*) indicates the best result.

Isoleucinomycin

| | Space group $P1$ ($n$ = 336) | | Space group $P2_12_12_1$ ($n$ = 84) | |
|---|---|---|---|---|
| Minimal function | Success rate (%) | Cost effectiveness (solutions h$^{-1}$) | Success rate | Cost effectiveness (solutions h$^{-1}$) |
| PS(90°, 2) | 8.5 | 6.38 | 10.4 | 20.8 |
| COS | 8.8 | 5.21 | 10.2 | 21.5 |
| EXP | 17.4* | 8.72 | 16.3 | 33.9* |

Crambin

| | Space group $P1$ ($n$ = 654) | | Space group $P2_1$ ($n$ = 327) | |
|---|---|---|---|---|
| Minimal function | Success rate (%) | Cost effectiveness (solutions h$^{-1}$) | Success rate (%) | Cost effectiveness (solutions h$^{-1}$) |
| PS(90°, 2) | 0.6 | 0.08 | 4.8 | 0.78 |
| COS | 8.4 | 0.82 | 5.2 | 0.81 |
| EXP | 32.2* | 2.86* | 5.6 | 0.92 |

known for space groups with higher symmetry. However, even for other space groups, *Exponential Shake-and-Bake* has the potential to outperform *Shake-and-Bake* with the default double 90° shift as implemented in an earlier version of *SnB* (*i.e.* v1.5) (Miller *et al.*, 1994). It seems clear that the radius of convergence of the exponential minimal function is larger than that of the cosine minimal function. Preliminary experiments, using truncated data for the alpha-1 structure at 1.1 Å, indicate that the exponential function does not significantly alter the lower-resolution limit for *Shake-and-Bake* applications.

The precise reason for the high success rate in space group $P1$ obtained with *Exponential Shake-and-Bake* remains unknown. It appears that the parameter-shift size $S$ is a critical parameter for very large $P1$ structures. For instance, the success rate for triclinic lysozyme obtained with the exponential minimal function varies from 0.0 to 40.0% when the parameter-shift size $S$ changes by only 10°. This observation suggests that the range of parameter-shift values yielding significant success rates may decrease rapidly as the number of non-H atoms in the unit cell increases. Therefore, it may be advisable to vary the shift angle for large structures using a relatively fine grid in order to avoid missing solutions altogether.

The applications of the exponential minimal function to structures in space groups $P2_1$ and $P2_12_12_1$ are complicated by the problem of predicting good parameter values. Nevertheless, the results presented in Table 8 indicate that structures in space group $P2_1$ can be efficiently treated as if they were $P2_1$ structures, by employing the exponential minimal function with default parameters given by equations (32)–(34). Treating $P2_12_12_1$ structures in space group $P1$ will improve the success rate but, due to the fourfold increase in computational effort, the computational efficiency will be reduced.

We would like to express our appreciation to Thomas Tan, Andrew Polumbo and Jan Pevzner for assisting in the development of *SnB* v2.0. We also thank the Center for Computational Science and Technology (CCST), Mathematics and Computer Science Division, Argonne National Laboratory, and the Cornell Theory Center (CTC) at Cornell University for allowing us to use their High Performance Computing Research Facilities. This research was supported by NIH grant GM-46733 and NSF grant IRI-9412415.

## References

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

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

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

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

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

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

Hauptman, H. A. (1991). *Crystallographic Computing 5: from Chemistry to Biology*, edited by D. Moras, A. D. Podnarny & J. C. Thierry, pp. 324–332. IUCr/Oxford University Press.

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.

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

Langs, D. A. (1988). *Science*, **241**, 188–191.

Loll, P. J., Bevivino, A. E., Korty, B. D. & Axelsen, P. H. (1997). *J. Am. Chem. Soc.* **119**, 1516–1522.

Marshall, G. R., Hodgkin, E. E., Langs, D. A., Smith, G. D., Zabrocki, J. & Leplawy, M. T. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 487–491.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.

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

Pletnev, V. Z., Galitskii, N. M., Smith, G. D., Weeks, C. M. & Duax, W. L. (1980). *Biopolymers*, **19**, 1517–1534.

Pletnev, V. Z., Ivanov, V. T., Langs, D. A., Strong, P. & Duax, W. L. (1992). *Biopolymers*, **32**, 819–827.

Privé, G. G., Anderson, D. H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). *Protein Sci.* **8**, 1–9.

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.

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.

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

Weeks, C. M., Hauptman, H. A., Chang, C.-S. & Miller, R. (1994). *Likelihood, Bayesian, Inference and their Application to the Solution of New Structures. Am. Crystallogr. Assoc. Trans.*, Vol. 30, edited by G. Bricogne & C. W. Carter, pp. 153–161.

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

Weeks, C. M. & Miller, R. (1999*b*). *Acta Cryst.* D**55**, 492–500.