

# SnB: Applying Shake-and-Bake to Proteins

Charles M. Weeks

Mol. Biophys. Dept., Hauptman-Woodward MRI, 73 High Street, Buffalo, NY 14203-1196, USA  
weeks@hwi.buffalo.edu

Russ Miller

Mol. Biophys. Dept., Hauptman-Woodward MRI, 73 High Street, Buffalo, NY 14203-1196, USA  
Dept. of Comp. Sci., SUNY-Buffalo, Buffalo, NY 14260, USA  
miller@hwi.buffalo.edu

## Abstract

The Shake-and-Bake algorithm is a powerful formulation of direct methods which alternates reciprocal-space phase refinement with filtering in real space to impose constraints. As implemented in SnB version 1.5, the current distributed version of the computer program, Shake-and-Bake combines peak picking in real space with optimization via either parameter-shift reduction of the minimal-function value or tangent-formula refinement. The procedure employs a multisolution approach in which initial trial structures consist of randomly positioned atoms. The SnB program has provided *ab initio* solutions for protein structures containing as many as 600 independent non-H atoms, provided that good-quality diffraction data are available to 1.1Å resolution.

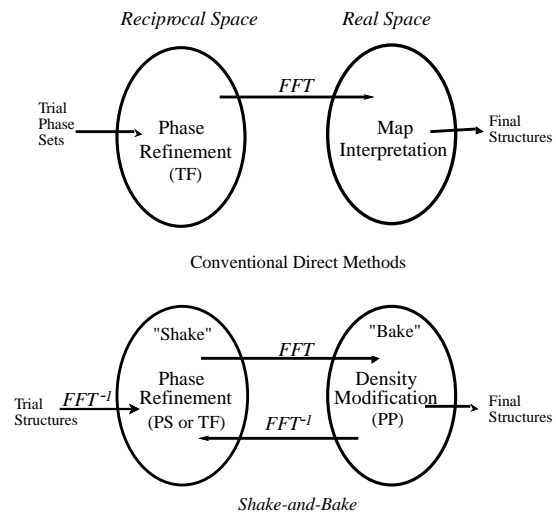
## 1 Introduction

The successful application of conventional direct methods to the *ab initio* solution of structures large enough to be regarded as small proteins has been the exclusive province of expert practitioners. Such noteworthy applications to structures in the 300-atom range include avian pancreatic polypeptide [1] and gramicidin A [2]. The development of the *Shake-and-Bake* algorithm [3-5] and its implementation in the computer program *SnB* [6] has made feasible the routine solution of structures of this size. *SnB* has been widely distributed and used with default parameters in other laboratories to solve structures containing as many as 450 atoms.

In the conventional direct methods approach, multiple sets of trial phases are refined using the tangent formula [7]. Several iterations (passes through the phase list) are made, and the final phase sets are then ranked according to figures-of-merit. One or more of the most promising combinations are then transformed to real space and, if possible, the corresponding electron density maps are interpreted in terms of atomic structures. The quality of a basically correct model structure may be significantly improved by doing a few cycles of Fourier refinement, a process which Sheldrick [8] has termed E-Fourier recycling. Another form of recycling was introduced by Jerome Karle [9] who recognized that even a

relatively small, chemically sensible fragment extracted by manual interpretation of an E-map could be parlayed into a complete solution by transformation back to reciprocal space and then performing additional iterations of tangent-formula refinement.

The tremendous increases in computer speed in recent years have made it feasible to consider cycling every trial structure back-and-forth between real and reciprocal space many times, while performing optimization alternately in each space. This is a compute-intensive task, as it requires the use of two Fourier transforms during each cycle. This cyclical process forms the basis of the synergistic *Shake* (phase refinement) and *Bake* (density modification) procedure in which the power of reciprocal-space phase refinement is augmented by filtering to impose the phase constraints implicit in real space. The *Shake-and-Bake* algorithm is compared to the conventional procedure in Figure 1.



**Figure 1. A comparison of conventional direct methods with Shake-and-Bake.**

TF = tangent formula; PS = parameter shift;  
PP = peak picking.

In *SnB* version 1.5, phases are assigned initial values by generating trial structures consisting of randomly positioned atoms (thereby avoiding overly consistent phase sets) and then computing structure factors. The percentage of such trial structures that converge to solution is a function of, among other things, size and complexity of the structure, resolution and quality of data, and space group, as well as the number of refinement cycles. As one might expect, for structures in a given space group, the success rate typically decreases as the size of the structure increases. Success rates for structures in P1 are significantly higher than for other space groups. This may be related to the fact that the origin position may be chosen arbitrarily in P1.

Automatic real-space electron-density map interpretation consists of selecting an appropriate number of the largest peaks (typically equal to or less than the expected number of atoms) to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance between atoms. If markedly unequal atoms are present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space. Thus, *a priori* knowledge concerning the chemical composition of the crystal is utilized, but no knowledge of constitution is required or used during peak selection. It is useful to think of peak picking in this context as simply an extreme form of density modification appropriate when atomic-resolution data are available. The entire dual-space refinement procedure is repeated for an appropriate number of cycles which have been determined empirically by experimentation with known datasets [5].

### 1.1 The Minimal Function

Most applications of *Shake-and-Bake* have also differed from conventional direct methods in that the phase-refinement portion of the cycle has been based on a simple parameter-shift procedure [5] which reduces the value of the minimal function,

$$\begin{aligned}
R(\phi) = & \left( \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \{ \cos(\phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}}) \right. \\
& - \left. [I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})] \right\}^2 + \sum_{\mathbf{L},\mathbf{M},\mathbf{N}} |B_{\mathbf{LMN}}| \\
& \times \left\{ \cos(\phi_{\mathbf{L}} + \phi_{\mathbf{M}} + \phi_{\mathbf{N}} + \phi_{-\mathbf{L}-\mathbf{M}-\mathbf{N}}) \right. \\
& - \left. [I_1(B_{\mathbf{LMN}})/I_0(B_{\mathbf{LMN}})] \right\}^2 \Big)^{-1} \\
& \times \left( \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} + \sum_{\mathbf{L},\mathbf{M},\mathbf{N}} |B_{\mathbf{LMN}}| \right)^{-1}
\end{aligned} \tag{1}$$

[10-12]. The minimal function expresses a relationship among phases related by triplet and negative quartet invariants which have the associated parameters (or weights)

$$A_{\mathbf{HK}} = \left( \frac{2}{N^{1/2}} \right) |E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{H}+\mathbf{K}}|, \tag{2}$$

and

$$\begin{aligned}
B_{\mathbf{LMN}} = & (2/N) |E_{\mathbf{L}} E_{\mathbf{M}} E_{\mathbf{N}} E_{\mathbf{L}+\mathbf{M}+\mathbf{N}}| \left[ \left( |E_{\mathbf{L}+\mathbf{M}}|^2 \right. \right. \\
& \left. \left. + |E_{\mathbf{M}+\mathbf{N}}|^2 + |E_{\mathbf{N}+\mathbf{L}}|^2 \right) - 2 \right],
\end{aligned} \tag{3}$$

respectively, where the  $|E|$ 's are the normalized structure factor magnitudes and  $N$  is the number of atoms, assumed identical, in the unit cell.  $R(\phi)$  is a measure of the mean square difference between the calculated structure invariants and their expected values as given by the ratio of Bessel functions, and it is expected to have a minimum,  $R_T$ , when the phases are equal to their correct values for some choice of origin and enantiomorph. The formula for  $R_T$ ,

$$\begin{aligned}
R_T = & \frac{1}{2} + \left( \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} + \sum_{\mathbf{L},\mathbf{M},\mathbf{N}} |B_{\mathbf{LMN}}| \right)^{-1} \\
& \times \left[ \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \left\{ \left( \frac{[I_2(A_{\mathbf{HK}})]}{[2I_0(A_{\mathbf{HK}})]} \right) \right. \right. \\
& - \left. \left. \left\{ \frac{[I_1(A_{\mathbf{HK}})]}{[I_0(A_{\mathbf{HK}})]} \right\}^2 \right) \right. \right. \\
& + \left. \sum_{\mathbf{L},\mathbf{M},\mathbf{N}} |B_{\mathbf{LMN}}| \left\{ \left( \frac{[I_2(B_{\mathbf{LMN}})]}{[2I_0(B_{\mathbf{LMN}})]} \right) \right. \right. \\
& - \left. \left. \left\{ \frac{[I_1(B_{\mathbf{LMN}})]}{[I_0(B_{\mathbf{LMN}})]} \right\}^2 \right) \right. \left. \right] < \frac{1}{2},
\end{aligned} \tag{4}$$

does not require prior knowledge of the phases and therefore can be calculated *ab initio*. Experimentation has thus far confirmed that: (i) the minimal function,

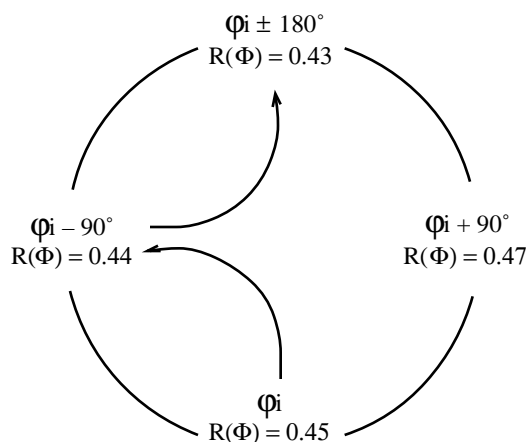
when used actively in the phasing process, is diagnostic in that a histogram of  $R(\phi)$  values for the refined trial structures can be used with high confidence to decide whether or not a solution exists and, (ii) when solutions do exist, the final trial structure corresponding to the smallest value of  $R(\phi)$  is a solution.

## 1.2 Phase Refinement

Parameter shift is a seemingly simple search technique that has proven to be quite powerful as an optimization method when used in conjunction with the minimal function, provided that appropriate choices of parameter values are made. The phases are considered in decreasing order with respect to the values of the associated  $|E|$ 's. When considering a given phase  $\phi_i$ , as shown in Figure 2, the value of the minimal function is initially evaluated three times. First, with the given set of phase assignments, second with phase  $\phi_i$  modified by the addition of the predetermined phase shift, and third with  $\phi_i$  modified by the subtraction of the predetermined phase shift. If the first evaluation yields the minimum of these three values of the minimal function, then consideration of  $\phi_i$  is complete, and parameter shift proceeds to  $\phi_{i+1}$ . Otherwise, the direction of search is determined by the modification that yields the minimum value, and the phase is updated to reflect that modification. In this case, phase  $\phi_i$  continues to be updated by the predetermined phase shift in the direction just determined so long as the value of the minimal function is reduced, though there is a user-defined predetermined maximum number of times that the shift is attempted. Based on extensive experimentation with these and related parameters, involving a variety of structures in several space groups, it has been determined that in terms of running time and percentage of trial structures that produce a solution, an excellent choice of parameters consists of the following: (i) perform a single pass through the phase set, (ii) evaluate the phases in order by decreasing  $|E|$ -values, and (iii) for each phase, perform a maximum of two  $90^\circ$  phase shifts [5].

When the parameter-shift phase refinement is applied in centrosymmetric space groups, only a single shift of  $180^\circ$  is required for each phase. Theoretically, it would seem as if restricted phases in acentric space groups should be handled in a similar fashion. In practice, however, this turns out not to be the case, at least in the space group  $P2_12_12_1$ . Higher success rates have been obtained in this space group if all phases are treated as general phases.

The traditional tangent-formula-based phase refinement of conventional direct methods has also been substituted for parameter-shift phase refinement in *Shake-and-Bake* and compared using known atomic-resolution datasets [13]. In this situation, the minimal



**Figure 2. An example of parameter shift with a maximum of two  $90^\circ$  phase shifts.**

Notice that initially the minimal function is calculated with the current set of phases, yielding a value of 0.45. The minimal function is then re-evaluated at  $\phi_i+90^\circ$  and  $\phi_i-90^\circ$ , yielding values of 0.47 and 0.44, respectively. Therefore,  $\phi_i$  is updated to  $\phi_i-90^\circ$ . Finally, the minimal function is evaluated once more by subtracting another  $90^\circ$  from  $\phi_i$ , which yields 0.43, and  $\phi_i$  is updated appropriately.

function is also computed, but used only as a figure-of-merit. Regardless of which refinement method is used, optimization proceeds most rapidly when there is immediate feedback of each refined phase value. In general, the tangent formula solves small structures (<100 atoms) more cost-effectively, but the two phase-refinement methods are equally efficient for solving most of the tested structures with more than 100 independent atoms, including crambin [14,15]. However, only the minimal function has produced recognizable solutions for gramicidin A. Approximately 5000 gramicidin A trial structures have been processed by each optimization method, and the minimal function has yielded 12 solutions (success rate of 0.25%). The tangent formula has, in fact, produced one solution, but this solution would not have been recognized if gramicidin A were an unknown because it had a relatively high value for the minimal function. This suggests that the minimal function is not a suitable figure-of-merit when it is used passively to trace the progress of tangent-formula phasing.

Tangent-formula cost-effectiveness is highly dependent on the number of phase-refinement iterations (*i.e.*, the number of passes through the list of phases) per complete *Shake-and-Bake* cycle whereas the minimal function does not exhibit such strong dependency. The number of tangent-formula iterations

per cycle must be chosen judiciously if high efficiency is, in fact, to be achieved. This is especially true for structures in space group P1 where it is never advisable to perform more than one iteration of tangent refinement per cycle. For example, the success rates of a 74-atom emerimycin peptide fragment [16] and a 96-atom enkephalin analog [17] drop from 57% to 4% and from 30% to 2%, respectively, when the number of iterations per cycle is increased from 1 to 2.

## 2 Methods

The *SnB* program has been described in the *Journal of Applied Crystallography* [6] and in the *User's Manual for Version 1.5.0* [18]. *SnB* is written in a combination of C and Fortran. Fundamental crystallographic routines are in Fortran, but C was chosen as a front-end language to facilitate the development of a user-friendly interface, dynamic allocation of memory, and the spawning of processes. There is a home page for *SnB* on the World Wide Web at URL: <http://www.hwi.buffalo.edu/SnB>; this home page is directly accessible from the ACA home page. Fundamental information is provided including a brief description of the procedure, a list of personnel, critical citations, announcements, bug reports/fixes, a manual corresponding to the current distributed version, and general information on how to obtain a copy of the program. *SnB* has been incorporated into Molecular Structure Corporation's *teXsan* package of crystallographic programs, and supercomputer versions have been installed on the Cray T3D and Cray C90 at the Pittsburgh Supercomputing Center, the CM-5 at NCSA, and the SP2 at the Cornell Theory Center. Stand-alone UNIX versions for SGI, SUN, IBM, and DEC alpha workstations are available, as are PC/Linux versions, directly from the Hauptman-Woodward Medical Research Institute. Interested persons should send an email message to [snb-requests@hwi.buffalo.edu](mailto:snb-requests@hwi.buffalo.edu).

### 2.1 Overview of the *SnB* Program

There are three major components of the *SnB* program. The first component performs the actual *Shake-and-Bake* structure-determination procedure by generating and processing trial structures. The second component permits the user to examine interactively the progress of a previously submitted structure-determination procedure. This component produces a histogram of the final  $R(\phi)$  values for all processed trial structures from which the user can decide whether or not a probable solution has been obtained. Finally, the third component permits the user to examine the geometry of the current best (lowest  $R(\phi)$ ) trial structure.

The main menu, shown in Figure 3, gives the user the basic options of (i) attempting to process trial structures to solve a structure, (ii) producing a

histogram of  $R(\phi)$  values for completed trial structures of a previously submitted structure-determination process, and (iii) displaying the best current structure for a previously submitted structure-determination process. It also permits the user to (iv) list the currently active structure-determination processes, or (v) exit from the program. A typical application of *SnB* consists of submitting a structure-determination process, monitoring the progress of the trial structures by occasionally viewing a histogram of final minimal-function values and, when a potential solution is identified, examining the geometry of this structure. The running time of the structure-determination procedure for large, difficult structures requiring many trials is substantial, and the ability to follow conveniently the course of such jobs is essential.

SnB

Crystal Structure Determination by Shake-and-Bake

COPYRIGHT 1993 by Russ Miller and  
Charles M. Weeks

MAIN MENU

1. Initiate Shake-and-Bake on trial structures.
2. Produce a histogram of completed trial structures.
3. Display the current best trial structure.
4. List active Shake-and-Bake jobs.
5. Exit.

Please enter your selection:

**Figure 3. The main menu of SnB.**

The flow chart presented in Figure 4 illustrates the basic operation of the *Shake-and-Bake* process. Triplet and (optionally) negative-quartet structure invariants, as well as the initial coordinates for the trial structures, must be generated. Once this information has been obtained, every trial structure is subjected to the following *Shake-and-Bake* procedure. Initially, a structure-factor calculation is performed which yields phases corresponding to the trial structure. The associated value of the minimal function,  $R(\phi)$ , is then computed. At this point, the cyclical *Shake-and-Bake* phasing procedure is initiated, as follows. The phases are refined *via* the tangent formula or by parameter shift so as to reduce the value of  $R(\phi)$ . These phases are then passed to a Fourier routine which produces an electron-density map, but no graphical output is produced. Instead, the map is examined by a peak-picking routine which typically finds the  $n$  largest peaks (where  $n$  is the number of independent non-H atoms in the asymmetric unit) subject to the constraint that no two peaks are closer than a specified distance.

These peaks are then considered to be atoms, and the process of structure-factor calculation, phase refinement, and density modification *via* peak selection is repeated for the predetermined number of *Shake-and-Bake* cycles.

For each completed trial structure, the final value of the minimal function is stored in a file which is subsequently used for histogramming purposes. In addition, a separate file is maintained which allows the user to examine the geometry of the best final structure. This file, which is updated at the completion of every trial structure, contains the final minimal function value as well as the initial and final peak or atom coordinates associated with the best trial (*i.e.*, the lowest  $R(\phi)$  value) processed so far. In *SnB* version 1.5, each trial is processed sequentially to completion. In the future, it is hoped that criteria permitting the early termination of unsuccessful trials can be incorporated.

## 2.2 Program Operation

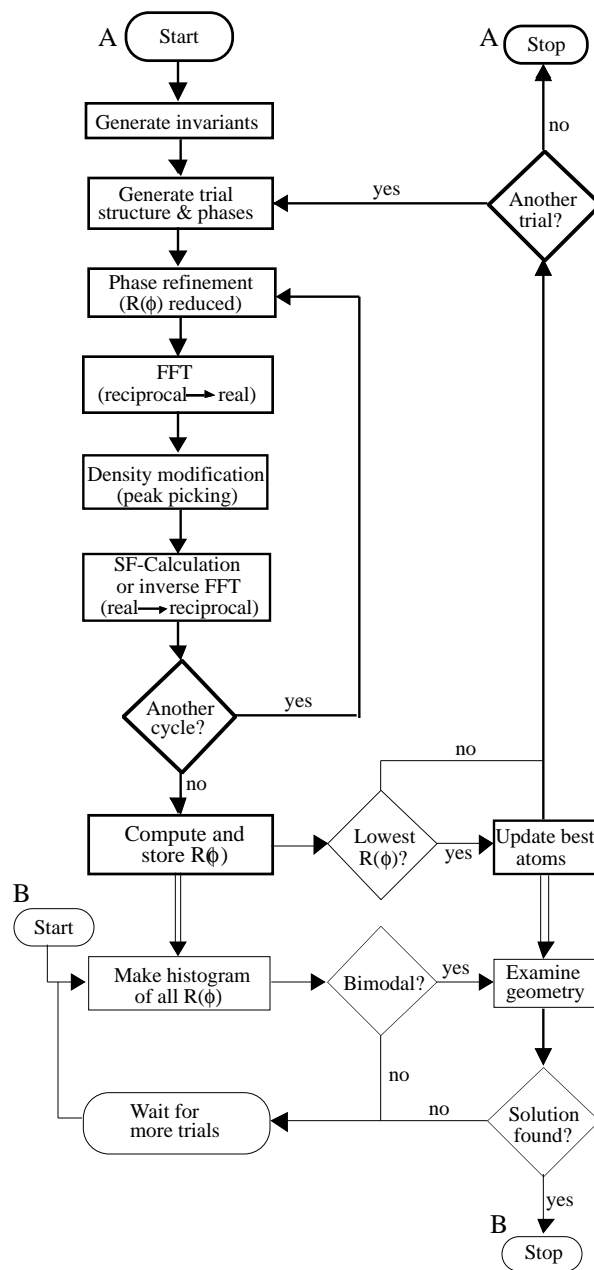
The current version of *SnB* interactively queries the user for a variety of information. Default values (displayed in square brackets following the query) are provided by the system for all critical parameters except the information specific for an individual structure (*e.g.*, cell constants). In addition, the user must supply an input reflection file consisting of  $h$ ,  $k$ ,  $l$  and the normalized structure-factor magnitudes,  $|E|$ . The program will automatically sort this data into descending order by  $|E|$ , eliminate systematic absences, and eliminate duplicate reflections. No selection based on  $\sigma(F)$  or  $F/\sigma(F)$  is performed. It is often critical that  $|E|$  values be calculated extremely carefully. Blessing's suite of programs [19] is recommended for this purpose.

### Structure-Determination Procedure.

Two modes of operation, novice and expert, are provided. The user is initially asked to provide a structure ID, which will be used as a file prefix for the structure under consideration. He or she is then prompted for some basic crystal data (space group, cell constants, and the contents of the asymmetric unit), as well as values for the parameters which control the course of *Shake-and-Bake*. The user operating in novice mode only needs to select the number of phases and invariants, specify the number of trials to be generated and processed, and choose the number of *Shake-and-Bake* cycles. The user operating in expert mode has more flexibility, including the use of alternative phase-refinement procedures.

Cost-effective default values for the control parameters are based on experience with several known test structures and are summarized in Table 1. Several parameters, including the numbers of phases and invariants to be used, depend on structure size and can

be expressed as a function of  $n$ . In general, inclusion of negative quartets in the invariant set improves the success rate but usually not in a cost-effective manner. Consequently, the default condition is to omit the negative quartets.



**Figure 4. A flow chart for the Shake-and-Bake algorithm.** Single lines represent flow of control; double lines show movement of data. 'Start A' represents the beginning of a structure-determination process, and 'Start B' indicates the beginning of a session in which the  $R(\phi)$  histogram and molecular geometry are checked.

**Table 1. Default parameter values for the SnB structure-determination procedure.**

<u>Parameter</u>	<u>Default</u>
Non-H atoms in asymmetric unit	$n$
Invariant generation	
Number of phases	$10n$
Number of triples	$100n$
Number of negative quartets	0
Starting atoms per random trial	$\min(n, 100)$
Number of <i>SnB</i> cycles	
Parameter shift (PS) refinement or Tangent formula refinement	$n/2$ $n/4$
PS phase refinement	
Size of phase shift	$90^\circ$
Maximum number of shifts	2
Number of iterations	1
Exploit restricted phases?	No
Number of peaks to select	$[0.8n, n]$
Exploit heavy atoms?	Yes
Number E-Fourier recycling steps	2

In order to generate an initial set of phases for each trial structure, the *Shake-and-Bake* method employs a structure-factor calculation based on initial trial structures or models. *SnB* can either generate a set of initial trial structures containing randomly positioned atoms or obtain a set of trial structures from the user. In practice, it is not necessary to use more than 100 randomly positioned atoms as a trial structure. Experimentation has shown that, during later cycles, choosing  $n$  peaks to recycle through the procedure gives optimum success rates for smaller structures. However, for large structures that are likely to contain a significant number of atoms with low occupancy or high thermal motion unlikely to be discernible in electron-density maps unless the phases are extremely accurate, trial structures composed of less than  $n$  peaks (*e.g.*,  $0.8*n$ ) give better performance. In the situation where trial structures are being generated by *SnB*, an initial seed is requested for use with the random-number generator that positions the atoms in each trial structure. It should be noted that the seed is solicited for the purpose of reproducibility of results.

Tests with several known data sets have focused on determining the cycle during which trial structures converge to solution. Notice that given a fixed number of machine cycles, it is important to consider the trade-off between the number of trial

structures processed and the number of cycles processed per trial structure. This experimentation has shown that, with a phase-refinement technique consisting of a single-iteration, two-step parameter shift of  $90^\circ$ , the point of diminishing returns is at approximately  $n/2$  cycles. Therefore, the program defaults the number of cycles per trial to approximately this value.

When the structure under consideration consists solely of atoms with atomic numbers less than 10, the program considers all atoms to be of equal weight for purposes of the structure-factor calculations. However, when atoms with atomic numbers greater than 10 are present, the user has the option of considering the appropriate number of largest peaks to be weighted by such values, though all atoms with atomic number less than 10 will be assigned a weight of 6. This use of information concerning the presence of heavier atoms to provide unequal weighting has resulted in accelerated convergence to solution in the case of structures containing a small amount of sulfur, iron, or chlorine atoms.

The final parameters to be chosen are concerned with E-Fourier recycling. These include the number of Fourier refinement cycles (*i.e.*, the number of *SnB* cycles with no phase refinement) and the number of peaks to select in each of these cycles. In the case of larger structures, it is useful to build, over the course of several cycles, from the number of peaks used during the *Shake-and-Bake* stage to the approximate total number of atoms expected in the structure.

After the dialogue is complete, the user is asked to review the information supplied and make any necessary changes, as illustrated in Figure 5 for a 64-residue scorpion toxin, Tox II. This information is then stored for use at a later time and for use by the histogram routine. Once a user decides that the set of parameters is satisfactory, the program automatically initiates the *Shake-and-Bake* structure-determination procedure by spawning a batch job.

**Histogram Procedure.** The histogram routine is supplied so that the user can easily determine whether or not a solution appears to be present in the set of completed trial structures. This routine supplies the user with a list of available results from previous and current structure-determination runs. After choosing one, the user is queried for the desired number of histogram buckets based on final minimal function ( $R(\phi)$ ) values. A bimodal distribution with significant separation is a typical indication that solutions are present (as shown in Figure 6), while a unimodal, bell-shaped distribution (*e.g.*, Figure 6 with the '0.467 to 0.470' row omitted) typically indicates a set of nonsolutions.

1. Search path: ./
2. Structure ID: ToxII
3. Space group: P212121
4. Cell constants:
  - A: 45.90 ALPHA: 90.00
  - B: 40.70 BETA : 90.00
  - C: 30.10 GAMMA: 90.00
5. Contents of the asymmetric unit: C500,S8
6. Generate new invariant set: Yes
  - Number of phases to use: 5000
  - Number of triples to use: 50000
  - Number of negative quartets to use: 0
  - Save invariants to file: ./ToxII.inv
7. Generate random trial structures: Yes
  - Number of trials to generate: 2000
  - Random number seed: 11909
  - Minimum interatomic distance: 1.20
  - Starting atoms per trial: 200
  - Save random trials to file: ./ToxII.random\_trials
8. Trial processing information
  - Number of trials to process: 2000
  - Beginning at trial number: 1
  - Number of Shake-and-Bake cycles: 255
9. Exploit knowledge of heavy atoms: Yes
10. Refinement method: Parameter Shift
  - Exploit knowledge of restricted phases: No
  - Number of complete passes through phase set: 1
  - Number of attempted phase shifts per pass: 2
  - Attempted phase shift per pass:
    - Pass #1: 90
11. Number of peaks to select: 400
12. Number of E-Fourier filtering cycles: 5
  - Number of peaks picked in cycle #1: 400
  - Number of peaks picked in cycle #2: 425
  - Number of peaks picked in cycle #3: 450
  - Number of peaks picked in cycle #4: 475
  - Number of peaks picked in cycle #5: 500

Would you like to make any changes? (y/n)

**Figure 5. Reviewing the values entered for the crystal data and control parameters.** The interatomic distances (item 7) and the information in items 9 - 12 are displayed and may be changed only when operating in expert mode.

**Geometric Examination.** The user is provided with two options for viewing the current best structure. The first requires only a character-based terminal and produces a text plot suitable for printing on a line printer. The user can then manually 'connect the dots.' This routine also produces a list of the interpeak distances and angles. The second option makes use of *GeomView*, a graphical routine developed by the Geometry Center and suitable for an X-Windows environment [20]. A binary version of *GeomView* is distributed with *SnB*. *GeomView* can also be obtained

by anonymous ftp at ftp.geom.umn.edu or on the World Wide Web at http://www.geom.umn.edu.

These options are included to assist the user in deciding whether a solution has, in fact, been obtained. They are not intended to provide complete visualization, especially for larger structures. The coordinates are available in a file and can be input into other graphical programs for more extensive display.

### 3 Results

The *SnB* program has been used to determine numerous structures in a variety of space groups. A list of successful applications to protein structures is given in Table 2. Gramicidin A, crambin, and rubredoxin were previously known test structures resolved at the Hauptman-Woodward Institute. The 64-residue scorpion toxin (Tox II) had been previously solved, but the number of residues and the amino acid sequence were deliberately withheld from the Buffalo group. The only information supplied (by Steve Ealick's group at CHESS) was that the protein was

Structure Name: ToxII	
Number of Atoms: 508	Number SnB Cycles: 255
Number of trials: 1619	Number of Phases: 5000
Lowest R( $\phi$ ): 0.467	Number of Triples: 50000
Highest R( $\phi$ ): 0.532	Number of Quartets: 0

R( $\phi$ ) Range	Trials	in range
0.467 to 0.470	1	*
0.471 to 0.474	0	
0.475 to 0.478	0	
0.479 to 0.482	0	
0.483 to 0.486	0	
0.487 to 0.490	0	
0.491 to 0.494	0	
0.495 to 0.498	0	
0.499 to 0.502	0	
0.503 to 0.506	0	
0.507 to 0.510	25	**
0.511 to 0.514	135	*****
0.515 to 0.518	386	*****
0.519 to 0.522	639	*****
0.523 to 0.526	390	*****
0.527 to 0.530	41	**
0.531 to 0.534	2	*
0.535 to 0.538	0	
0.539 to 0.542	0	
0.543 to 0.546	0	

**Figure 6. A 20-bucket histogram of the final minimal function values after 255 cycles for the 624-atom Tox II structure.** The separation between the single solution and the 1618 non-solutions is clearly shown.

**Table 2. Protein structures solved ab initio by SnB.**

<u>Structure</u>	<u>Non-H Atoms</u>	<u>Space Group</u>	<u>Resolution</u>	<u>Success Rate</u>	<u>References</u>
Vancomycin	255	P4 <sub>3</sub> 2 <sub>1</sub> 2	0.9Å	1/4200	[21]
Gramicidin A	317	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	0.86	0.25%	[2,22]
Er-1 Pheromone	325	C2	1.0	0.25%	[23]
Crambin	400	P2 <sub>1</sub>	0.83	2-3%	[14,15]
Alpha-1 Peptide	450	P1	0.92	5%	[24]
Rubredoxin	500	P2 <sub>1</sub>	1.0	2.7%	[22,25]
Tox II	624	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	0.96	1/1619	[26]

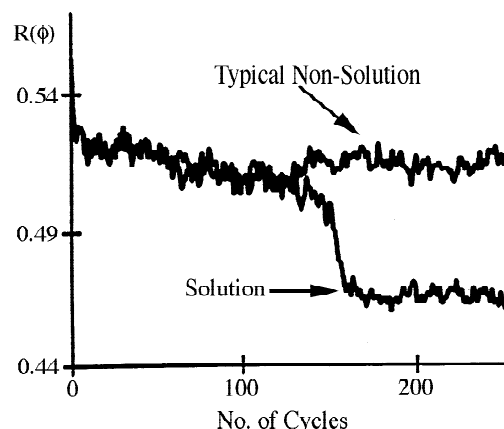
composed of approximately 500 atoms and contained four disulfide bonds. The remaining structures (vancomycin [21], Er-1 pheromone [23], and alpha-1 peptide [24]) were previously unknown, and the applications were made in other laboratories without direct involvement by the authors of *SnB*. All were solved routinely and automatically using essentially default parameters.

The application to Tox II was made on a network of SGI R4000 Indigo Workstations with *SnB* running as a background job for approximately six weeks. One morning, the histogram reproduced in Figure 6 was found during the daily progress check. After detecting that the histogram was now bimodal, the single trial in the 0.467 to 0.470 range was examined, and a conservative model consisting of five fragments and a total of 241 atoms was constructed. Following multiple cycles of Xplor refinement, the residual was 0.16 for 624 non-H atoms [26]. Figure 7 shows the course of the minimal function  $R(\phi)$ , as a function of cycle number, for the trial which led to the solution and for a typical non-solution trial. Both trials show almost identical behavior for approximately 130 cycles. Notice that  $R(\phi)$  for the trial that went to solution then drops precipitously from a value of about 0.50 to 0.467 and remains at that level for all remaining cycles. For the non-solution trial, however,  $R(\phi)$  oscillates between 0.51 and 0.52 for all remaining cycles [27].

It has been known for some time that conventional direct methods can be a valuable tool for locating the positions of heavy atoms using isomorphous  $\Delta E$ 's [28] and anomalous scatterers using anomalous  $\Delta E$ 's [29]. Thus, it is no surprise that the *Shake-and-Bake* algorithm can be fruitfully applied in this arena as well. The first application of this type was to native and Se-Met data for avian sarcoma virus integrase [30]. The four Se atoms were found using 189  $\Delta E$  values ( $>1.76$ ) in the resolution range 20 to 3.7Å. The investigators report that the isomorphous difference Patterson map was impossible to deconvolute without the aid of direct methods.

## 4 Concluding Remarks

The *SnB* program is currently undergoing major revisions. *SnB* version 2, targeted for release in late 1997, not only expands the capabilities currently available in *SnB* version 1.5, but will also significantly improve the running time of the procedure. The calculation of normalized structure-factor magnitudes ( $|E|$ 's) will be included, as will a more convenient interface to map interpretation programs. It should be noted that the percentage of time spent in the structure factor calculation is a function of the size of the structure. That is, for larger structures, a higher percentage of the time is spent in the structure-factor routine. The prototype *SnB* version 2 currently includes an inverse FFT which is much more efficient than structure-factor calculation for protein-sized molecules. In addition, the use of the inverse transform opens the door to density-modification protocols other than peak picking. Such protocols are likely to increase the range of applicability of the *Shake-and-Bake* method. The scope of the method can also be enlarged through consideration of invariant values individually estimated using SIR [31] or SAS [32] information and appropriate objective functions such as the SAS maximal function or tangent formula [33]. It



**Figure 7. Traces of  $R(\phi)$  for Tox II, as a function of cycle number, for the solution trial and for a typical non-solution trial.**



should also be noted that the *Shake-and-Bake* algorithm, with tangent-formula phase refinement only, has also been combined with iterative peaklist optimization [34].

The ultimate potential of the *Shake-and-Bake* approach to the *ab initio* structure determination of macromolecules is unknown. The combination of this technique with increasingly powerful computers has recently permitted direct-method solutions in situations regarded as impossible only a few years ago. The combination of *Shake-and-Bake* methodology with alternative density-modification methods and supplemental phasing information from isomorphous replacement and single- or multiple-wavelength anomalous dispersion may allow equally spectacular advances in the near future.

## Acknowledgments

The *Shake-and-Bake* algorithm and the *SnB* program have been made possible by the financial support of grants GM-46733 from NIH and IRI-9412415 from NSF. The authors would also like to acknowledge the guidance and inspiration provided by Prof. Herbert Hauptman throughout the development of *SnB*. Our sincere thanks are also given to the students (Chun-Shi Chang, Steven Gallo, Hanif Khalak, Jan Pevzner, and Pamela Thuman) whose labors have helped to make *SnB* a reality.

## References

- [1] M. M. Woolfson & J.-X. Yao, "On the Application of Phase Relationships to Complex Structures. XXX. *Ab Initio* Solution of a Small Protein by SAYTAN", *Acta Cryst.* **A46**, 409-413, 1990.
- [2] D. A. Langs, "Three-Dimensional Structure at 0.86 Å of the Uncomplexed Form of the Transmembrane Ion Channel Peptide Gramicidin A", *Science*, **241**, 188-191, 1988.
- [3] C. M. Weeks, G. T. DeTitta, R. Miller, & H. A. Hauptman, "Applications of the Minimal Principle to Peptide Structures", *Acta Cryst.* **D49**, 179-181, 1993.
- [4] R. Miller, G. T. DeTitta, R. Jones, D. A. Langs, C. M. Weeks, & H. A. Hauptman, "On the Application of the Minimal Principle to Solve Unknown Structures", *Science*, **259**, 1430-1433, 1993.
- [5] C. M. Weeks, G. T. DeTitta, H. A. Hauptman, P. Thuman, & R. Miller, "Structure Solution by Minimal Function Phase Refinement and Fourier Filtering: II. Implementation and Applications", *Acta Cryst.* **A50**, 210-220, 1994.
- [6] R. Miller, S. M. Gallo, H. G. Khalak, & C. M. Weeks, "*SnB*: Crystal Structure Determination via *Shake-and-Bake*", *J. Appl. Cryst.* **27**, 613-621, 1994.
- [7] J. Karle & H. Hauptman, "A Theory of Phase Determination for the Four Types of Non-Centrosymmetric Space Groups 1P22, 2P22, 3P12, 3P22", *Acta Cryst.* **9**, 635-651, 1956.
- [8] G. M. Sheldrick, "SHELX-84", in *Crystallographic Computing 3: Data Collection, Structure Determination, Proteins, and Databases*, G. M. Sheldrick, C. Kruger & R. Goddard (Eds.), Clarendon Press, Oxford, 1985, pp. 184-189.
- [9] J. Karle, "Partial Structural Information Combined with the Tangent Formula for Noncentrosymmetric Crystals", *Acta Cryst.* **B24**, 182-186, 1968.
- [10] T. Debaerdemaeker & M. M. Woolfson, "On the Application of Phase Relationships to Complex Structures. XXII. Techniques for Random Phase Refinement", *Acta Cryst.* **A39**, 193-196, 1983.
- [11] H. A. Hauptman, "A Minimal Principle in the Phase Problem", in *Crystallographic Computing 5: From Chemistry to Biology*, D. Moras, A. D. Podnarny & J. C. Thierry (Eds.), IUCr Oxford Univ. Press, 1991, pp. 324-332.
- [12] G. T. DeTitta, C. M. Weeks, P. Thuman, R. Miller, & H. A. Hauptman, "Structure Solution by Minimal Function Phase Refinement and Fourier Filtering: I. Theoretical Basis", *Acta Cryst.* **A50**, 203-210, 1994.
- [13] C. M. Weeks, H. A. Hauptman, C.-S. Chang, & R. Miller, "Structure Determination by *Shake-and-Bake* with Tangent Refinement", *ACA Transactions Symposium*, Vol. **30**, 1994.
- [14] W. A. Hendrickson & M. M. Teeter, "Structure of the Hydrophobic Protein Crambin Determined Directly from the Anomalous Scattering of Sulfur", *Nature*, **290**, 107-113, 1981.
- [15] C. M. Weeks, H. A. Hauptman, G. D. Smith, R. H. Blessing, M. M. Teeter, & R. Miller, "Crambin: A Direct Solution for a 400 Atom Structure", *Acta Cryst.* **D51**, 33-38, 1995.
- [16] G. R. Marshall, E. E. Hodgkin, D. A. Langs, G. D. Smith, J. Zabrocki, & M. T. Leplawy, "Factors Governing Helical Preference of Peptides Containing Multiple  $\alpha,\alpha$ -Dialkyl Amino Acids", *Proc. Natl. Acad. Sci. USA*, **87**, 487-491, 1990.
- [17] J. L. Krstenansky, D. A. Langs, & G. D. Smith, unpublished.
- [18] R. Miller, S. M. Gallo, H. G. Khalak, & C. M. Weeks, "*SnB*: A Structure Determination Package, User's Manual for Version 1.5.0".
- [19] R. H. Blessing, D. Y. Guo, & D. A. Langs, "Statistical Expectation Value of the Debye-Waller Factor and  $E(hkl)$  Values for Macromolecular Crystals", *Acta Cryst.* **D52**, 257-266, 1996.
- [20] *GeomView* was developed by the Geometry Center (Center for the Computation and Visualization of Geometric Structures) at the University of Minnesota.
- [21] P. J. Loll, personal communication.
- [22] H. A. Hauptman, "Looking Ahead", *Acta Cryst.* **B51**, 416-422, 1995.
- [23] D. H. Anderson, M. S. Weiss, & D. Eisenberg, "A Challenging Case for Protein Crystal Structure Determination: the Mating Pheromone Er-1 from *Euplotes raikovi*", *Acta Cryst.* **D52**, 469-480, 1996.
- [24] G. Prive, N. Ogihara, L. Wesson, D. Cascio, & D. Eisenberg, "A Designer Peptide at High Resolution:

- Shake-and-Bake* Solution of a 400 Atom Structure", Abstract W008, Proc. of the Am. Crystallogr. Assoc. Meeting, Montreal, 1995.
- [25] Z. Dauter, L. C. Sieker, & K. S. Wilson, "Refinement of Rubredoxin from *Desulfovibrio vulgaris* at 1.0Å with and without Restraints", *Acta Cryst.* **B48**, 42-59, 1992.
- [26] G. D. Smith, R. H. Blessing, S. E. Ealick, J. C. Fontecilla-Camps, H. A. Hauptman, D. Housset, D. A. Langs, & R. Miller, "The *ab Initio* Structure and Refinement of a Scorpion Protein Toxin", Abstract MS02.06.07, IUCr Meeting, Seattle, WA (USA), 1996.
- [27] H. A. Hauptman, "*Shake-and-Bake*, an Algorithm for the Automatic Solution, *ab initio*, of Crystal Structures", in *Macromolecular Crystallography*, Vol. II, ed. by C. W. Carter & R. Sweet, in press.
- [28] K. S. Wilson, "The Application of MULTAN to the Analysis of Isomorphous Derivatives in Protein Crystallography", *Acta Cryst.* **B34**, 1599-1608, 1978.
- [29] A. K. Mukherjee, J. R. Helliwell, & P. Main, "The Use of MULTAN to Locate the Positions of Anomalous Scatterers", *Acta Cryst.* **A45**, 715-718, 1989.
- [30] G. Bujacz, M. Jaskolski, J. Alexandratos, A. Wlodawer, G. Merkel, R. A. Katz, & A. M. Skalka, "High-Resolution Structure of the Catalytic Domain of Avian Sarcoma Virus Integrase", *J. Mol. Biol.* **253**, 333-346, 1995.
- [31] H. A. Hauptman, "On Integrating the Techniques of Direct Methods and Isomorphous Replacement I. The Theoretical Basis", *Acta Cryst.* **A38**, 289-294, 1982.
- [32] H. A. Hauptman, "On Integrating the Techniques of Direct Methods with Anomalous Dispersion. I. The Theoretical Basis", *Acta Cryst.* **A38**, 632-641, 1982.
- [33] H. A. Hauptman, "The SAS Maximal Principle: a New Approach to the Phase Problem", *Acta Cryst.* **A52**, 490-496, 1996.
- [34] G. M. Sheldrick & R. O. Gould, "Structure Solution by Iterative Peaklist Optimization and Tangent Expansion in Space Group P1", *Acta Cryst.* **B51**, 423-431, 1995.