J. Appl. Cryst. (1994). **27**, 613–621

SnB: Crystal Structure Determination *via* Shake-and-Bake

BY RUSS MILLER AND STEVEN M. GALLO

Department of Computer Science, State University of New York at Buffalo, Buffalo, NY 14260, *USA*

AND HANIF G. KHALAK AND CHARLES M. WEEKS

The Medical Foundation of Buffalo, 73 *High Street, Buffalo, NY* 14203, *USA*

(*Received* 8 *October* 1993; *accepted* 5 *January* 1994)

Abstract

Shake-and-bake is a direct-methods phasing algorithm for structure determination based on the *minimal principle*. *SnB* is a program based on shake-and-bake that has been used successfully to solve more than a dozen structures in a variety of space groups. The focus of this paper is on the details of this program, including its structure, system requirements, running times and the rationale for coding in a combination of C and Fortran. A summary of successful *SnB* applications is also provided. These include solving two previously unknown 100-atom structures and re-solving crambin (a structure containing the equivalent of approximately 400 fully occupied atomic positions) for the first time with a direct-methods technique.

1. Introduction

The shake-and-bake method of crystal structure determination is a direct-methods phasing algorithm (Weeks, DeTitta, Miller & Hauptman, 1993; Miller, DeTitta, Jones, Langs, Weeks & Hauptman, 1993; Weeks, DeTitta, Hauptman, Thuman & Miller, 1994) based on the minimal principle (Hauptman, 1988, 1991; DeTitta, Weeks, Thuman, Miller & Hauptman, 1994). (Readers unfamiliar with the minimal principle are encouraged to review the Appendix.) As illustrated in Fig. 1, the shake-and-bake method alternates phase refinement in reciprocal space with density modification in real space in an attempt to reach the global minimum of the minimal function, $R(\varphi)$. As with traditional direct-methods techniques, numerous trial structures or sets of phases are evaluated. Unlike traditional direct methods, however, reciprocal-space refinement requires reducing the value of $R(\varphi)$ (*e.g. via* a parameter-shift algorithm), and an essential feature of the process is the repeated automatic alternation between reciprocal and real space. Experimentation has thus far confirmed that: (i) the minimal function is diagnostic in that a histogram of such values corresponding to the resultant structures can be used with high confidence for the purpose of determining whether or not a solution exists; and (ii) when solutions do exist, the final structure corresponding to the smallest value of the minimal function is a solution.

This paper describes a computer program, *SnB*, which implements the shake-and-bake method. §2 gives an overview of the structure of the program, including the coordination of the critical routines. §3 gives details concerning the phase-refinement techniques that are currently available in *SnB*. §4 discusses the user interface, including the crystallographic input required, default values, and options for evaluating potential solutions. §5 focuses on the files that are manipulated during execution of the program and §6 discusses language and operating system requirements. Finally, examples of successful applications are given in §7.

2. Overview of the *SnB* program

The structure diagram presented in Fig. 2 illustrates the basic organization of the *SnB* program. There are three major components. 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 $R(\varphi)$ values for all processed trial structures from which the user can decide whether or not a probable solution has been obtained.

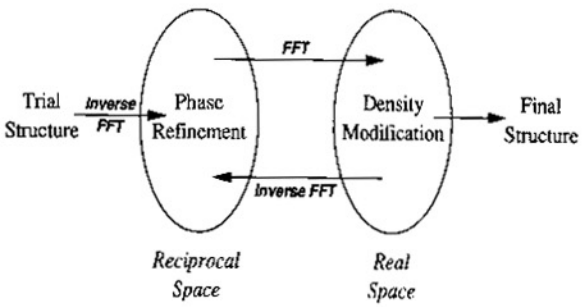

Fig. 1. The shake-and-bake method of structure determination alternates between real and reciprocal space. In the current program, phase refinement is by parameter shift and density modification consists of peak picking.

The third component permits the user to examine the geometry of the current best trial structure.

Pseudo-code for the structure-determination component of the package is given in Fig. 3. Notice that, in order to perform the aforementioned shake-and-bake process, triplet and 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 that yields phases corresponding to the trial structure. The associated value of the minimal function, $R(\varphi)$, is then computed. At this point, the cyclical shake-and-bake phasing procedure is initiated, as follows. The phases are refined so as to reduce the value of $R(\varphi)$. These phases are then passed to a Fourier routine that produces an electron-density map. No graphical output is produced. Instead, the map is then examined by a peak-picking routine that finds the n largest peaks (for an n-atom structure) 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.

Typical running times for the overall shake-and-bake process and its major steps are presented in Table 1. A space-group-general fast Fourier transform is used to compute electron-density maps, but conventional methods are employed in the structure-factor summation. The later calculation takes a large percentage of the total computation time, especially for large structures. It is anticipated that the present routine will be replaced by a space-group-general inverse Fourier transformation in a future version of the program. In addition, space-group-specific routines for the most common space groups, such as $P2_1$ and $P2_12_12_1$, will also be incorporated.

For each completed trial structure, the final value of the minimal function is stored in a file that is subsequently used for histogramming purposes. In addition, a separate file is maintained that allows the user to

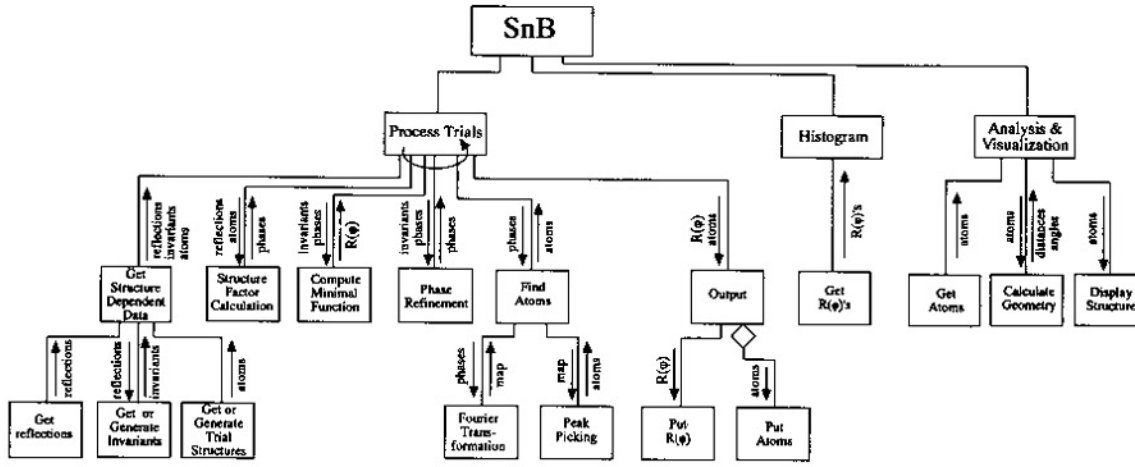

Fig. 2. A structure diagram for the *SnB* program.

```
Generate triplet and (optionally) negative quartet invariants
Generate trial structures with randomly positioned atoms
For every trial structure do
  Structure-factor calculation {produce phase set φ}
    Compute the initial value of the minimal function, R(φ)
    Do for a specified number of shake-and-bake cycles
      Phase refinement {modify phase set φ}
      Fourier transformation {produce electron density map}
      Peak selection {select largest peaks as new set of atoms}
      Structure-factor calculation {produce revised phase set φ}
      Compute new R(φ)
    end Do
    Store final R(φ) for histogram purposes
    If R(φ) < R_best
      R_best = R(φ)
      Store atoms
    end If
end For
```

Fig. 3. Pseudo-code for the structure-determination routine. A histogram of $R(\varphi)$ values and the geometry of the current best structure can be displayed at any time after the process has been initiated.

Table 1. *Sample running times for SnB*

Times are given in seconds per shake-and-bake cycle per trial per Sparc processor on a CM-5.

Structure (atoms)	Structure factor	Minimal function	Phase refinement	Fourier peak picking	Total
9α-Methoxycortisol (28)	0.5	0.06	0.4	7.5	8.5
Emerimycin (74)	0.9	0.1	1.1	8.1	10.2
Isoleucinomycin (84)	4.4	0.2	1.3	18.2	24.1
Crambin (400)	47.3	0.7	8.3	34.7	91.0

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 $R(\varphi)$ value] processed so far. In the present version of *SnB*, 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.

3. Phase refinement

The current version of *SnB* refines phases by either parameter shift or a related global binary search technique. Other refinement methods under investigation include gradient descent and simulated annealing.

Parameter shift is a seemingly simple search technique that has proven to be quite powerful when 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 a given phase φ_i is considered, as shown in Fig. 4, the value of the minimal function is initially evaluated three times. First, with the given set of phase assignments, second, with phase φ_i modified by the addition of the predetermined phase shift, and, third, with φ_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 φ_i is complete and parameter shift proceeds to φ_{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 φ_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 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:

1. perform a single pass through the phase set;
2. evaluate the phases in order of decreasing $|E|$ values;
3. for each phase, perform a maximum of two 90° phase shifts.

The global binary search is a multiple-pass single-shift variant of parameter shift in which the shift size is halved in each subsequent pass through the phase set. Five passes through the phase set, where the first pass has a phase shift of either 90 or 120°, produces a relatively high number of successful trials. Notice that such schemes require several more evaluations of the minimal function than the two-step one-iteration method. Further, experimentation has shown that, in most cases, the percentage of trial structures that yield a solution is inferior to the optimized parameter-shift method just described.

When the parameter-shift phase refinement is applied in centrosymmetric space groups, only a single shift of 180° is required for each phase. Theoretically, it would seem as if restricted phases in non-centrosymmetric 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.

4. Program operation

The current version of *SnB* operates interactively by querying the user for a variety of information. Default values (displayed in square brackets following the query)

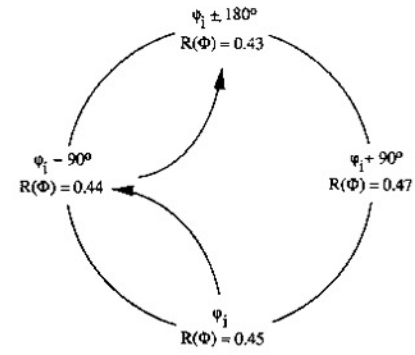

Fig. 4. An example of parameter shift with a maximum of two 90° phase shifts. Notice that initially the minimal function is calculated with the current set of phases, yielding a value of 0.45. The minimal function is then re-evaluated at $\varphi_i+90°$ and $\varphi_i-90°$, yielding values of 0.47 and 0.44, respectively. Therefore, φ_i is updated to $\varphi_i-90°$. Finally, the minimal function is evaluated once more by subtracting another 90° from φ_i, which yields 0.43, and φ_i is updated appropriately

are provided by the system for all critical parameters except the basic structure-dependent crystallographic information. 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 these 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.

In this paper, we restrict our attention to the character-terminal-based implementation of *SnB*. The main menu, shown in Fig. 5, gives the user the basic options of (i) attempting to process trial structures to solve a structure, (ii) producing a histogram of $R(\varphi)$ 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) to exit 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 then, when a potential solution is identified, examining the geometry of this structure.

4.1. *Structure-determination procedure*

The dialog associated with the structure-determination procedure is outlined in Fig. 6. After deciding whether to operate in novice or expert mode, the user is initially asked to provide a structure ID, which will be used as a file prefix for the structure under consideration. The user 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 that 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 alternative phase-refinement procedures.

Cost-effective default values for the control parameters are based on experience with several known test

```
                              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:
```

Fig. 5. The main menu of *SnB*.

Table 2. *Default values for SnB parameters*

The number of starting atoms per trial depends on the space group.

Parameter	Default value
Independent non-H atoms in asymmetric unit	n
Invariant generation	
Number of phases	$10*n$
Number of triplets	$100*n$
Number of negative quartets	0
Random trial generation	
Minimum interatomic distance	1.2 Å
Minimum intermolecular distance	2.7 Å
Minimum next-nearest-neighbor distance	2.0 Å
Number of starting atoms per trial	1, 2, or 4
Number of *SnB* cycles	$\sim 0.5*n$
Structure-factor calculation	
Exploit knowledge of heavy atoms?	Yes
Phase refinement	
Number of iterations	1
Size of phase shift	90°
Maximum number of phase shifts	2
Exploit knowledge of restricted phases?	No
Fourier transformation	
Number of peaks to select	n

structures and are summarized in Table 2. Several parameters depend on structure size and can be expressed as a function of n, the number of independent non-H atoms in the asymmetric unit. In general, inclusion of negative quartets in the invariant set improves the success rate

```
Interactive setup
  Do you want to use information from a previously stored structure [no]:
  Search path for data (reflection/invariant/trials) files [.]:
  File prefix to the structure you want to process [ ]: pr435
  Confirm that pr435 is what you want [yes]: yes

    Input crystal data
      Enter space group [P212121]:
      Enter the cell constants in decimal form. Give angles in degrees.
      A      : 13.1
      B      : 14.075
      C      : 10.658
      ALPHA  : 90.
      BETA   : 90.
      GAMMA: 90.
      Contents of the asymmetric unit [e.g. O18.N6,C60,H102]: O6,C22,H32

    Input SnB parameter values
      Generate new invariant set [no]:
      Enter the number of phases to be used [280]:
      Enter the number of triples to be used [2800]:
      Enter the number of negative quartets to be used [0]:
      Generate random trial structures [yes]: no
      Enter name of input atom file [./pr435.trials]:
      Enter initial structure number [1]:
      Number of trials to process [64]: 500
      Enter the number of shake-and-bake cycles to be performed [100]: 10

  Would you like to make any changes? (y/n) n
  Would you like to save this information to a file? (y/n) y
  Enter the complete path and name of the file: pr435.info.novice
  Would you like to run the program **now** with the parameters defined
    as above? (y/n) y

Submit spawned process
  Do you wish to continue a previous run [no]:
  Please enter a prefix name for the files that will contain
    the results of this run [May30.1807]:

  Running SnB with file prefix 'May30.1807'...
```

Fig. 6. Dialog for the structure-determination procedure (novice mode). User responses are in *italics*.

but often not in a cost-effective manner. Consequently, the default condition is to omit the negative quartets but their use should be considered, especially in space groups lacking a screw axis or glide plane if a run with default values does not produce a solution.

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. Experimentation has shown that the initial structure should contain a sufficient number of atoms and their symmetry-related mates so as to specify the origin and enantiomorph with respect to the relevant space group. In the case that *SnB* is used to create a set of trial structures, the default number of atoms generated for each trial structure is the minimum number of atoms needed to specify the origin and enantiomorph, but this number can be as large as n, if desired. In addition, for 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. A variety of other values are also solicited for the purpose of generating chemically sound sets of randomly positioned atoms to be used as trial structures. The default *minimum interatomic distance* between atoms in a generated trial structure (all symmetry-related positions considered) is 1.2 Å. Further, there can be no more than one distance (a possible intramolecular distance) between any pair of independent atoms or their symmetry equivalents less than the value of the minimum intramolecular distance. Distances less than the next-nearest-neighbor distance are considered to be bonds, and no atom is permitted to have more than four such distances.

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. The experimentation has shown that, with a phase refinement technique consisting of a single-iteration two-step parameter shift of 90°, 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, although 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. Finally, the user may choose to treat less than n peaks as atoms in the structure-factor calculations if it is expected that some atoms will be difficult or impossible to locate during the early stages of the phasing process because they may have low occupancy or high thermal motion.

After the basic dialog is complete, the user is asked to review the information supplied and make any necessary changes, as illustrated in Fig. 7 for the 28-atom test structure, 9α-methoxycortisol (structure ID = pr435). This information may then be stored for use at a later time or for use by the histogram routine. Once a user decides that the set of parameters is satisfactory then in order to initiate the corresponding shake-and-bake process the user is queried for a prefix that will be used for the files that will be created during execution. These files will be discussed in detail in §5.

4.2. *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

```
1. Search path: ./
2. Structure ID: pr435
3. Space group: P212121
4. Cell constants:
      A: 13.1600 ALPHA: 90.0000
      B: 14.0760 BETA : 90.0000
      C: 10.6560 GAMMA: 90.0000
5. Contents of the asymmetric unit: C6/O2/N1
6. Generate new invariant set: Yes
      Number of phases to use: 280
      Number of triples to use: 3820
      Number of negative quartets to use: 0
      Save invariants to file: ./pr435.inv
7. Generate random trial structures: Yes
      Number of trials to generate: 1000
      Random number seed: 13579
      Minimum interatomic distance: 1.20
      Minimum intermolecular distance: 2.70
      Minimum next nearest neighbor distance: 2.60
      Starting atoms per trial: 1
      Save random trials to file: ./pr435.random.trials
8. Trial processing information
      Number of trials to process: 500
      Beginning at trial number: 1
      Number of shake and bake cycles: 30
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: 28
12. Optional information storage
      Keep trace file containing Minimal Function values: Yes
      Store running times in file: Yes
      Store all final structures in file: Yes
Would you like to make any changes? (y/n)
```

Fig. 7. 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

structure-determination runs. After choosing one, the user is queried for the number of histogram buckets based on final minimal function (R_{min}) values. The user is also given the option of having the histogram appear on the screen or being recorded in a file. Fig. 8 shows a sample histogram in which the range of values of the minimal function has been divided into 20 buckets. A bimodal distribution with significant separation is a typical indication that solutions are present, while a unimodal bell-shaped distribution typically indicates a set of nonsolutions. A similar histogram is used in the program *SHELXS*86 (Sheldrick, 1986).

4.3. *Geometric examination*

Currently, only a rather crude 'visualization' tool for examining a potential solution is provided. This routine requires only a character-based terminal. It simply produces the distances and angles and provides a rough plot of the atoms/peaks. The user can then manually 'connect the dots'. It should be noted, however, that, since the set of coordinates is available in a file, this information may be input into any standard visualization package for a more graphical display.

5. File system

The present form of *SnB* requires an input file consisting of reflections and their associated $|E|$ values. *SnB* can then be used to generate an invariant file consisting of triples and quartets, as well as a file of initial trial structures. Alternatively, these files may be supplied by the user from either a previous *SnB* run or by some other means. After completing the structure determination dialog, the user has the ability to save the information provided to an information file. This information file can be used as a template for subsequent *SnB* runs or to provide some descriptive information for the histogram routine, as discussed below. When a structure-determination process begins to run, a file is written containing a verbose description of the basic information supplied by the user (*e.g.* the name of the structure, number of atoms, number of trials to process, number of invariants *etc.*). While processing the trials, the program creates a file consisting of the final value of the minimal function, $R(\varphi)$, for every trial. A best trial file is also maintained that contains the final $R(\varphi)$ value and associated coordinates of peaks (atoms) of the best trial structure [lowest $R(\varphi)$] processed to date, as well as the coordinates of the corresponding initial trial structure.

When the user requests a histogram, the necessary minimal-function values are read from the appropriate file and the resulting histogram is either displayed on the screen or sent to a file. This histogram also contains some of the descriptive information from the information file, as shown in Fig. 8. When the user requests a display of the best structure processed thus far, the set of peaks is read from the best trial file and the structure is processed and sent to a file for display purposes. A structure-determination process is physically initiated by spawning a Unix process to perform the shake-and-bake algorithm with the set of parameters and files defined. When the

```
                    Structure name: pr435
                  Prefix file name: pr435

   Number of atoms: 28                    Number of shake-and bake cycles: 30
   Number of trials processed: 500        Number of phases: 280
   Lowest Rmin value: 0.272               Number of triples: 2803
   Highest Rmin value: 0.582              Number of quartets: 0

Minimal function range     Trials in range

  0.272   to   0.287            64          ************************
  0.288   to   0.303             3          *
  0.304   to   0.319             1          *
  0.320   to   0.335             1          *
  0.336   to   0.351             0
  0.352   to   0.367             0
  0.368   to   0.383             0
  0.384   to   0.399             5          **
  0.400   to   0.415            20          *******
  0.416   to   0.431            20          *******
  0.432   to   0.447            41          ***************
  0.448   to   0.463            73          ***************************
  0.464   to   0.479           103          ***************************************
  0.480   to   0.495            82          ******************************
  0.496   to   0.511            54          ********************
  0.512   to   0.527            22          ********
  0.528   to   0.543             5          **
  0.544   to   0.559             4          *
  0.560   to   0.575             0
  0.576   to   0.591             2          *

 Hit RETURN to go back to the main menu.
```

Fig. 8. A sample histogram with 20 buckets. Note the bimodal distribution, which is indicative of a solution. In this case, there are 69 solutions.

Table 3. *Some successful applications of SnB*

These results were obtained using a one-pass, 90°, two-step parameter shift phase refinement and an atom:phase:triplet:negative-quartet ratio of 1:10:100:0. In each case, approximately 1000 trial structures were evaluated.

Structure	Space group	Atoms in asymmetric unit	*SnB* cycles	Success rate (%)
9α-Methoxycortisol	$P2_12_12_1$	28	30	14.5
Emerimycin	$P1$	74	100	91.4
Isoleucinomycin	$P2_12_12_1$	84	150	12.1
Meso-valinomycin	$P\bar{1}$	84	150	1.0
Non-peptidic enkephalin analog	$P1$	96	150	32.8
Ternatin II	$P2_12_12_1$	110	150	0.7
Cholesterol butanoate	$P2_1$	132	100	1.4
Valinomycin analog (1.1 Å)	$P2_1$	174	150	0.4
Valinomycin–dioxane	$P2_1$	176	150	2.1
Crambin	$P2_1$	~400	200	3.6

task is spawned, a file is created that contains the process ID of the task. This information is used to determine which processes are active at any given time.

There are several additional files that are maintained at the request of a user operating in expert mode. These include: (i) a trace file that records the value of the minimal function at the end of every shake-and-bake cycle for every trial structure; (ii) a file used to record the running times for the entire process as well as for each of the major routines; and (iii) a file that is used to record the final coordinates for every trial structure.

6. System requirements

Algorithm development and parameter tuning for *SnB* requires experimentation that is computationally intensive. Since this was clear from the inception, *SnB* was initially developed on massively parallel supercomputers, including the Intel iPSC/860 hypercube and Thinking Machines Corporation CM-2 and CM-5 systems. For a variety of mostly nontechnical reasons, the final stages of the development cycle concentrated on the CM-5 implementation. Once the CM-5 version of *SnB* became stable in terms of algorithmic development, parameter tuning and the user interface, the decision was made to retrofit the system for a workstation environment.

The initial workstation chosen as a target for *SnB* was a standard Sparc-based Sun workstation. This choice was made primarily for compatibility reasons in that both the CM-5 nodes and CM-5 front-end systems are Sparc based. We have since ported the system to SGI workstations. All of the previously mentioned platforms operate under the Unix operating system and some of the auxiliary routines, such as displaying active processes, exploit specific features of Unix. Further, our current implementation uses Unix directives to spawn a task when the user decides to initiate the structure-determination procedure. Currently, a task is either processed on the available workstation or on the back end (*i.e.* processing nodes) of a CM-5.

SnB is written in a combination of C and Fortran. Numerous fundamental crystallographic routines (Fourier routines, peak picking, structure-factor calculations *etc.*) and utilities had previously been written in Fortran. Therefore, in the interest of developing a system *via* rapid prototyping, it was decided to use Fortran for the numerically intensive routines. C was chosen as the front-end language for a variety of reasons, including (i) it promotes the development of a friendly user interface, (ii) it allows for dynamic allocation of memory, which provides for an efficient use of the *SnB* system in that only the amount of memory necessary for the structure under consideration will be appropriated and (iii) it facilitates the spawning of processes.

7. Structures determined with *SnB*

The *SnB* program has been used successfully to determine numerous structures in a variety of space groups. A list of some of these applications is given in Table 3. These structures range in size from a 25-atom prostaglandin structure in $P1$ to the $P2_1$ crambin structure with the equivalent of about 400 atoms. The ternatin structures were determined first by *SnB*. The other test structures had been previously solved by classical direct methods, although many required considerable painstaking non-routine efforts. All were solved quickly, routinely and automatically by *SnB*.

As one might expect, for structures in a given space group, the success rate typically decreases as the size of the structure increases. Also, notice that the success rates for structures in $P1$ are significantly higher than for other space groups. This may be related to the fact that there are an infinite number of choices of origin position in $P1$.

8. Summary and concluding remarks

In this paper, we discuss the basic structure of the *SnB* program, gave details of the phase-refinement options and discuss a variety of systems issues. It should be noted that a help facility is also available in *SnB*. The user simply enters 'help' as a response to any query

and is then given a help screen, which generally defines the terms used in the query and gives examples of responses for different situations. *SnB* has proved quite successful in determining structures larger than existing programs have been able to determine and in solving previously unknown structures that had escaped solution by traditional methods.

The package is presently available for Sun and SGI workstations. It is scheduled to be incorporated into packages from Molecular Structures Corporation and MacScience early in 1994. It is scheduled to be available on the Cray T3D and Cray C90 at the Pittsburgh Supercomputing Center by summer 1994. In addition, the program should be available on a Thinking Machines Corporation CM-5 early in 1994. Individuals wishing to obtain Unix workstation versions or a version for the CM-5, should contact the authors.

Future work consists of optimizing the individual pieces of the program, in particular, the Fourier, structure-factor and peak-picking routines, and considering tradeoffs in terms of fundamental parameters. For example, we have recently determined that the success rate will be improved if, for a structure consisting of n non-H atoms in the asymmetric unit, one uses n-atom trial structures as opposed to the minimal atom starts discussed in §4.1. The major scientific challenge we are now facing is that of extending the shake-and-bake method of structure determination to structures of lower resolution. At present, the peak-picking strategy seems to extend only to about 1.2 Å. We will use *SnB* as a test bed for replacing the peak-picking routine with a density-modification routine targeted at lower-resolution data.

We thank H. Hauptman for developing the minimal principle and M. Teeter and H. Hope for use of their crambin data. We thank C.-S. Chang, R. Jones, A. Khalak, S. Potter and P. Thuman for some of the early developmental work based on the shake-and-bake solution strategy. We also thank the Pittsburgh Supercomputing Center and Thinking Machines Corporation for allowing us to use their CM-5's, as well as the National Institutes of Health for allowing us to use their Intel iPSC/860 hypercube. This research grant was partially supported by NSF grant IRI-9108288 and NIH grant GM-46733.

APPENDIX
The minimal principle*

The minimal principle (Hauptman, 1991) is based on a new minimum-variance phase-invariant residual. Assume a crystal structure P to be fixed, but unknown *a priori*. The normalized structure-factor magnitudes $|E|$ are also assumed to be known. The function to be minimized, the so-called *minimal function*, is defined initially as a function, $R(\Phi)$, of the structure invariants $T_{\mathbf{HK}}$ and $Q_{\mathbf{LMN}}$,

$$R(\Phi) = \left\{ \sum_{\mathbf{H,K}} A_{\mathbf{HK}} \left[\cos T_{\mathbf{HK}} - \frac{I_1(A_{\mathbf{HK}})}{I_0(A_{\mathbf{HK}})} \right]^2 + \sum_{\mathbf{L,M,N}} |B_{\mathbf{LMN}}| \left[\cos Q_{\mathbf{LMN}} - \frac{I_1(B_{\mathbf{LMN}})}{I_0(B_{\mathbf{LMN}})} \right]^2 \right\} \times \left[\sum_{\mathbf{H,K}} A_{\mathbf{HK}} + \sum_{\mathbf{L,M,N}} |B_{\mathbf{LMN}}| \right]^{-1}. \quad (1)$$

We define a triple as

$$T_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \quad (2)$$

a quartet as

$$Q_{\mathbf{LMN}} = \varphi_{\mathbf{L}} + \varphi_{\mathbf{M}} + \varphi_{\mathbf{N}} + \varphi_{-\mathbf{L}-\mathbf{M}-\mathbf{N}}, \quad (3)$$

$$A_{\mathbf{HK}} = (2/N^{1/2})|E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{H+K}}|, \quad (4)$$

$$B_{\mathbf{LMN}} = (2/N)|E_{\mathbf{L}} E_{\mathbf{M}} E_{\mathbf{N}} E_{\mathbf{L+M+N}}| \times \left(|E_{\mathbf{L+M}}|^2 + |E_{\mathbf{M+N}}|^2 + |E_{\mathbf{N+L}}|^2 - 2 \right), \quad (5)$$

N to be the number of atoms, assumed identical, in the whole unit cell, and I_1 and I_0 to be modified Bessel functions. It should be noted that $B_{\mathbf{LMN}}$ [*cf.* (5)] can take negative values when the cross terms ($|E_{\mathbf{L-M}}|$, $|E_{\mathbf{M+N}}|$, $|E_{\mathbf{N+L}}|$) are very small, so when used as a weight in (1) its absolute value is taken. Further, it should be noted that the ratio of Bessel functions I_1/I_0 is known to be the expected value of the corresponding cosine. In view of (2) and (3), (1) also defines R as a function, $R(\varphi)$, of the phases. Because the magnitudes $|E|$ are presumed to be known, the functions $R(\Phi)$ and $R(\varphi)$ are well defined solely as functions of Φ and φ, respectively.

The phases are functions, for a fixed choice of origin and enantiomorph, of the atomic position vectors. Specifically,

$$E_{\mathbf{H}} = |E_{\mathbf{H}}| \exp(i\varphi_{\mathbf{H}}) = (1/N^{1/2}) \sum_{j=1}^{N} \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \quad (6)$$

where $\mathbf{r}_j$ is the position vector of the atom labeled j. Since the structure invariants $T_{\mathbf{HK}}$ and $Q_{\mathbf{LMN}}$ are

* This section appeared in essentially the same form in Miller, DeTitta, Jones, Langs, Weeks & Hauptman (1993).

uniquely determined for any given structure S, independent of the choice of origin, it follows that (1) also defines a function, $R(S)$, of structures S.

The minimal principle states that, among all phases φ that satisfy the necessary identities, those corresponding to the true structure P minimize $R(\varphi)$; or, equivalently, the (N-atom) structure S that minimizes $R(S)$ coincides with P.

References

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

Hauptman, H. A. (1988). Am. Crystallogr. Assoc. Meet., Philadelphia, USA, Abstract R4.

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

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

Sheldrick, G. M. (1986). *SHELXS86. Program for the Solution of Crystal Structures.* Univ. of Göttingen, Germany.

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

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