

An Introduction to *SnB* v2.0

Russ Miller^{a,b} & Charles M. Weeks^a

^aThe Hauptman-Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203

^bDepartment of Computer Science and Engineering, SUNY-Buffero, Buffalo NY 14260
{miller,weeks}@hwi.buffalo.edu; www.hwi.buffalo.edu/SnB/

Abstract. *SnB* is a computer program based on *Shake-and-Bake*, a direct-methods procedure for determining crystal structures. The program has been used in a routine fashion to solve difficult structures that could not be solved by traditional reciprocal-space routines based on the tangent formula. Recently, *SnB* has also been used to determine the Se sites in large selenomethionyl-substituted proteins. *SnB* v1.5 has been available for several years and is being used regularly in many laboratories. At this workshop, we introduce *SnB* v2.0, which incorporates a graphical user interface (GUI) written in Java, a dynamic histogram display, an integrated crystallographic data processing package, and an interactive Java/VRML-based visualization facility. In addition, *SnB* v2.0 provides the user with a variety of new algorithmic options.

I. INTRODUCTION

SnB is a publicly available direct-methods package based on the *Shake-and-Bake* method of structure determination. The program has been available since 1994 and has been available for download from the *SnB* Web site¹ since 1995. At the time of its introduction, tangent-based programs such as RANTAN and MULTAN were capable of routinely solving structures containing less than 100 nonhydrogen atoms and occasionally providing solutions for problems in the 100-200 atom range. Therefore, with its routine application to structures containing several hundred nonhydrogen atoms, *SnB* represented a significant advance in *ab initio* direct-methods phasing. In fact, due to the success of *SnB*, Sheldrick has recently exploited the *Shake-and-Bake* philosophy in a related “half-baked” (SHELXD) algorithm, which employs peaklist optimization. In addition to solving more complex structures than had previously been possible, *SnB* has also been used to increase the number of Se sites that can be located for selenomethionyl-substituted proteins. For example, *SnB* was used to initiate the structure determination process for 190kDa human placental S-adenosylhomocysteine (AdoHcy) hydrolase by finding the 30 Se atoms using peak anomalous difference data.

In Section II, we present an overview of direct methods, including the *Shake-and-Bake* procedure. In Section III, we give an overview of the *SnB* program, including features available in the current public release of the program, *SnB* v1.5, and details of implementation. We also discuss new features that are incorporated into *SnB* v2.0, including a graphical user interface (GUI), a graphical histogram display, an interactive visualization routine, optimizations to the *Shake-and-Bake* procedure, and an interface to the DREAR suite of data-processing programs. An appendix is included that contains examples of using *SnB* v2.0 for traditional single data sets, SAS, and SIR situations.

¹ www.hwi.buffalo.edu/SnB/

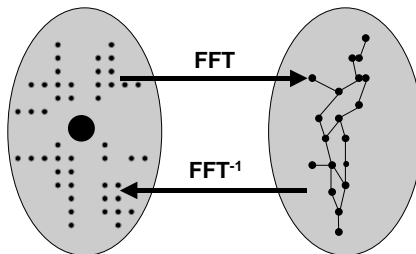


Figure 1. The relationship between reflections in reciprocal space (left oval) and the atoms in real space (right oval). Note that the locations and intensities of the reflections are measurable, but their phase values are not.

II. BACKGROUND

The tremendous increases in computer speed in recent years have made possible the development of a direct-methods multitrial or ‘multisolution’ technique [Germain & Woolfson, 1968] in which each trial structure is repeatedly cycled back-and-forth between real and reciprocal space, alternately performing optimization in each space, as shown in Figure 1. This compute-intensive process, which requires the use of two Fourier transforms during each cycle, is known as *Shake* (phase refinement) and *Bake* (density modification) [Miller et al., 1993; Weeks, DeTitta, Hauptman, Thuman, & Miller, 1994a]. This procedure has been described, in detail, in two recent reviews [Weeks & Miller, 1996; Weeks & Miller, 1997]. The ability to impose physically meaningful constraints in real space has increased the size of molecular structures amenable to phasing by direct methods from 100 to 1000 independent non-H atoms. The method known as iterative peaklist optimization [Sheldrick & Gould, 1995] has been patterned after *Shake-and-Bake* and relies even more heavily on real-space constraints.

Multitrial direct-methods procedures require multiple sets of starting phases, which can be subjected to a specified refinement protocol. In recent years, it has become routine to use a random number generator to assign initial phase values [Baggio, Woolfson, Declercq, & Germain, 1978; Yao, 1981]. In the *Shake-and-Bake* procedure, phases are assigned initial values by first generating trial structures consisting of randomly positioned atoms (thereby imposing an atomicity constraint from the outset) and then computing structure factors. The tangent formula

$$\tan(f_{\mathbf{H}}) = \frac{\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{\mathbf{H}-\mathbf{K}}| \sin(f_{\mathbf{K}} + f_{\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{\mathbf{H}-\mathbf{K}}| \cos(f_{\mathbf{K}} + f_{\mathbf{H}-\mathbf{K}})} \quad (1)$$

[Karle & Hauptman, 1956], in either its original or a weighted form [Hull & Irwin, 1978], provides the means for phase refinement in conventional multisolution phasing programs like MULTAN [Germain, Main, & Woolfson, 1971], RANTAN [Yao, 1981], and SHELXS [Sheldrick, 1985; Sheldrick, 1997].

On the other hand, *Shake-and-Bake* permits alternative optimization strategies during the phase-refinement step. In particular, an especially good strategy is to use parameter-shift search

[Bhuiya & Stanley, 1963] to reduce the value of an objective function such as the minimal function

$$R(f) = \frac{\sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \left\{ \cos(f_{\mathbf{H}} + f_{\mathbf{K}} + f_{-\mathbf{H}-\mathbf{K}}) - \frac{I_1(A_{\mathbf{HK}})}{I_0(A_{\mathbf{HK}})} \right\}^2}{\sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}} \quad (2)$$

[Debaerdemaeker & Woolfson, 1983; DeTitta, Weeks, Thuman, Miller, & Hauptman, 1994; Hauptman, 1991]. The minimal function expresses a relationship among phases related by triplet invariants that have the associated parameters (or weights)

$$A_{\mathbf{HK}} = \frac{2|E_{\mathbf{H}}E_{\mathbf{K}}E_{\mathbf{H}+\mathbf{K}}|}{N^{1/2}} \quad (3)$$

where the $|E|$'s are the normalized structure-factor magnitudes and N is the number of atoms, assumed identical, in the unit cell. The minimal function, $R(f)$, is a measure of the mean-square difference between the values of the triplet invariants calculated using a trial set of phases and their expected values (given by the ratio of modified Bessel functions, I_1 / I_0) as predicted by the conditional probability distribution of structure invariants [Cochran, 1955]. It is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph. Experimentation has thus far confirmed that *i*) when the minimal function is used actively in the *Shake-and-Bake* process and *ii*) solutions actually occur, the final trial structure corresponding to the smallest value of $R(f)$ is a solution. Therefore, $R(f)$ is also an extremely useful figure of merit for selecting those trials that have converged to solution.

In the applications reported to date, automatic real-space electron-density map interpretation in the *Shake-and-Bake* procedure consists of selecting an appropriate number of the largest peaks (equal to or less than the expected number of atoms in the structure) to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance. These peaks are then regarded as atoms, and a structure-factor calculation imposes the atomicity constraint. If markedly unequal atoms are known to be 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 a predetermined number of cycles or until it can be determined with high probability that the trial will not yield a solution.

Applications of *SnB*. Information is presented in Table 1 about a variety of protein structures that were solved by either *SnB* v1.5 or an alpha version of *SnB* v2.0. Gramicidin A [Hauptman, 1995], crambin [Weeks et al., 1995], rubredoxin [Hauptman, 1995], the 500-atom scorpion toxin II [Smith et al., 1997], and the 1000-atom lysozyme [Deacon, Weeks, Miller, & Ealick, 1998], were previously known test structures re-solved by the *SnB*. The remaining structures, including

PROTEIN STRUCTURE	LOCATION	ATOMS	SPACE GROUP	SUCCESS RATE
*Vancomycin (Tetragonal)	U. Penn	258	P4 ₃ 2 ₁ 2	0.8%
*Conotoxin EpI	HWI	289	I4	53.0
Gramicidin A	HWI	317	P2 ₁ 2 ₁ 2 ₁	1.1
*Er-1 pheromone	UCLA	328	C2	0.25
Crambin	HWI	~400	P2 ₁	4.8
*Alpha-1 peptide	OCI/Toronto	471	P1	5.0
Rubredoxin	HWI	497	P2 ₁	6.2
*Vancomycin (Triclinic)	HWI/U. Penn	547	P1	N.A.
Scorpion Toxin II	HWI	624	P2 ₁ 2 ₁ 2 ₁	1.4
Lysozyme	Cornell/HWI	~1200	P1	22.0

Table 1. A table of some successful *SnB* applications to proteins. The marked (*) structures were previously unknown. The number of atoms includes solvent molecules as well as protein. *Success rate* is the percentage of trial structures that go to solution.

STRUCTURE	LOCATION	SE ATOMS	PROTEIN SIZE (ASU)	SPACE GROUP
C3d	Toronto/HWI	8	34kDA	P2 ₁ 2 ₁ 2 ₁
GPATase	Purdue/HWI	22	112	C222 ₁
*AdoHcy	Toronto/HWI	30	95	C222

Table 2. A table of some successful *SnB* applications for determining the Se sites in selenomethionyl-substituted proteins. Several other proprietary Se-Met structures have been solved with *SnB*. The marked (*) structure was previously unknown.

the tetragonal form of vancomycin [Loll, Bevivino, Korty, & Axelsen, 1997], the triclinic form of vancomycin [Loll, Miller, Weeks, & Axelsen, 1998], conotoxin EpI [Hu et al., 1998], Er-1 pheromone [Anderson, Weiss, & Eisenberg, 1996], and alpha-1 peptide [Prive, Ogihara, Wesson, Cascio, & Eisenberg, 1995], were previously unknown. Furthermore, several of the applications to these previously unknown structures were made in other laboratories without direct involvement by the authors of *SnB*. The majority of the structures presented in Table 1 were solved routinely and automatically using default parameters. (Cost-effective default values are provided for all the control parameters based on extensive experimentation with several known test structures [Chang, Weeks, Miller, & Hauptman, 1997; Miller & Weeks, 1997; Weeks et al., 1994a; Weeks, Hauptman, Chang, & Miller, 1994b].) The success rates (*i.e.*, percentage of trial structures that go to solution) depend on size and complexity of the structure, resolution and quality of data, the presence of atoms heavier than oxygen, the space group, and the number of refinement cycles.

Se-Met Applications. An especially powerful procedure developed in recent years with the aid of tools from molecular biology involves the replacement of the naturally-occurring, sulfur-containing, amino acid residue methionine with the isomorphous residue selenomethionine (Se-Met), in which sulfur is replaced with the heavier element selenium (Se) [Doublet & Carter, 1992; Hendrickson et al., 1989]. It has been known for some time that conventional direct methods can be a valuable tool for locating the positions of heavy-atom substructures using

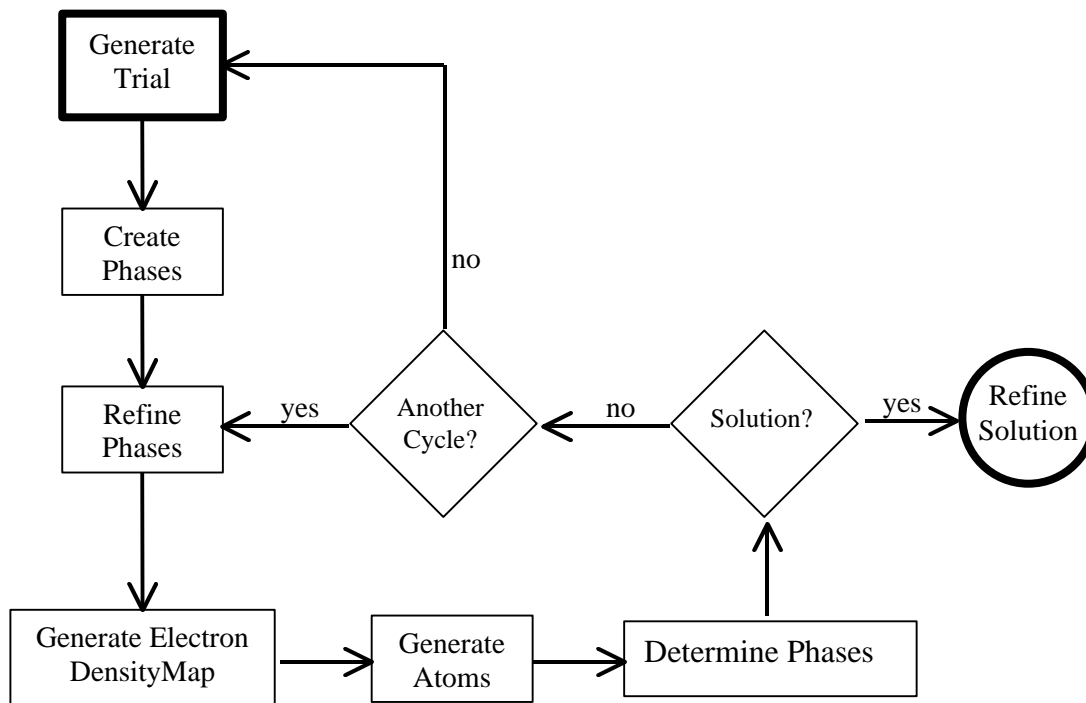


Figure 2. A generic presentation of the *Shake-and-Bake* procedure.

isomorphous [Wilson, 1978] and anomalous difference structure factors [Mukherjee, Helliwell, & Main, 1989]. *SnB* has recently been applied to several such selenomethionyl-substituted structures, as presented in Table 2. Highlights include the solution to a 190kDa human placental S-adenosylhomocysteine (AdHcy) hydrolase, which was initiated by exploiting *SnB* to determine the 30 Se atoms using peak anomalous difference data [Turner et al., 1998].

III. AN OVERVIEW OF *SNB*

SnB [Miller, Gallo, Khalak, & Weeks, 1994] is a user-friendly implementation of *Shake-and-Bake* that has been developed over the past 5 years. Pertinent information concerning *SnB* may be found at www.hwi.buffalo.edu/SnB. Stand-alone UNIX executables for SGI, SUN, IBM, and DEC alpha workstations, as well as PC/Linux versions, may be downloaded from this site. In addition, *SnB* has also been ported to a variety of supercomputers, including the Cray T3D/E, Cray C90, TCM CM-5, and IBM SP2. *SnB* is available in hundreds of laboratories worldwide.

A. *SnB* v1.5

The most recent public release of *SnB* is denoted as *SnB* v1.5. The main menu of *SnB* v1.5 gives the user the option of *i*) generating and processing trial structures in an effort to determine a structure by *Shake-and-Bake*, *ii*) producing a histogram of minimal function values corresponding to completed trial structures, and *iii*) displaying the best current trial structure. 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

user must supply *SnB* with *i*) basic information about the crystal (*e.g.*, its chemical contents) and *ii*) an input reflection file consisting of reciprocal-space positions and intensities (normalized structure-factor magnitudes, $|E|$). The program will automatically sort this data into descending order by $|E|$, eliminate systematic absences, and eliminate duplicate reflections.

Cost-effective default values for the control parameters (displayed following each query) are presented to the user, based on experience with several known test structures. A table of suggested values can also be found on the *SnB* Web site. The relative efficiency of tangent-formula and parameter-shift phase refinement in *Shake-and-Bake* has been compared using known atomic-resolution data sets [Chang et al., 1997]. In the case of tangent refinement, the minimal 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 parameter shift is more reliable for larger structures.

At the beginning of the structure determination procedure, a preprocessing step is performed that consists of generating structure invariants, as well as the initial (random) coordinates for trial structures. Once this information is available, every trial structure is subjected to the following procedure (refer to Figure 2).

1. Initially, a structure-factor calculation is performed that yields phases corresponding to the trial structure.
2. The associated value of the minimal function is then computed.
3. At this point, the cyclical *Shake-and-Bake* phasing procedure is initiated, as follows.
 - a) The phases are refined *via* the tangent formula or by parameter shift so as to reduce the value of the minimal function.
 - b) These phases are then passed to a Fourier routine that 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 expected number of independent nonhydrogen atoms), subject to the constraint that no two peaks are closer than a specified distance.
 - c) These peaks are then considered to be atoms.
 - d) A structure factor calculation is invoked in order to obtain phases corresponding to these atoms.
 - e) The process of phase refinement, density modification *via* peak selection, and structure-factor calculation 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, and the histogram routine can be run to determine whether or not a solution appears to be present in the set of completed trial structures. 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. Two options permit the user to view 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 (Center for the Computation and Visualization of

Geometric Structures at the University of Minnesota) that is suitable for an X-Windows environment. These options are included to assist the user in deciding whether a solution has, in fact, been obtained. The visualization routines provided in *SnB* v1.5 are not intended to support a complete analysis, especially for larger structures. It is expected that promising coordinates will be input into other graphical programs for more extensive display and refinement.

B. *SnB* v2.0

A completely redesigned version of *SnB* is targeted for beta-release during the Summer of 1998. This version, entitled *SnB* v2.0 [Weeks & Miller, 1998], is being constructed in an effort to

1. **improve the overall performance** of the program, so as to allow for the efficient determination of larger and more difficult structures,
2. provide the user with some fundamental crystallographic **data processing routines** that were missing from *SnB* v1.5 in order to automatically produce the input data files required by the *Shake-and-Bake* procedure,
3. provide a modern **graphical user interface**,
4. provide a **dynamic histogram facility** to aid in the diagnosis of solution,
5. provide the user with an improved **graphical visualization tool**, and
6. provide an easy means for porting the code to a **variety of platforms**, including workstations, PCs, NOWs, and multiprocessor supercomputers.

Programming Details. *SnB* consists of two major pieces of code, namely, the front-end interface and the back-end crystallographic package. The menu-driven, ASCII-based, front-end of *SnB* v1.5 was written in C, while its back-end was written in Fortran [Gallo, Miller, & Weeks, 1996]. *SnB* v2.0 includes a GUI front-end written in Java, and a significantly improved back-end, again written in Fortran. The core crystallographic routines were re-implemented from the ground up, which permitted a complete and thorough rethinking of the data structures in an effort to maximize efficiency. It should be noted that, when standard parameter settings are used for large structures, the new version of the program is significantly faster. *SnB* v1.5 provides only a structure-factor calculation for transforming from real to reciprocal space, whereas *SnB* v2.0 also includes an inverse FFT.

New Features. *SnB* v2.0 contains a graphical user interface (GUI) written in Java (Figure 3, right), a dynamic histogram display (Figure 3, left), and an interactive Java/VRML-based visualization facility (Figure 4). In addition, *SnB* v2.0 provides integrated access to data processing routines and provide the user with a variety of new algorithmic options. Descriptions of several of these features follow.

Integrated Data Processing. A major deficiency of *SnB* v1.5 was that it did not include a routine to generate $|E|_s$. This deficiency has been alleviated in *SnB* v2.0 by incorporating the DREAR package of data-processing routines [Blessing, Guo, & Langs, 1996]. This provides the user with the capability of automatically generating the $|E|_s$ that are required before invoking the *Shake-and-Bake* procedure. The interface provided with *SnB* v2.0 provides the user with the ability to process traditional single data sets, as well as SIR and SAS data sets.

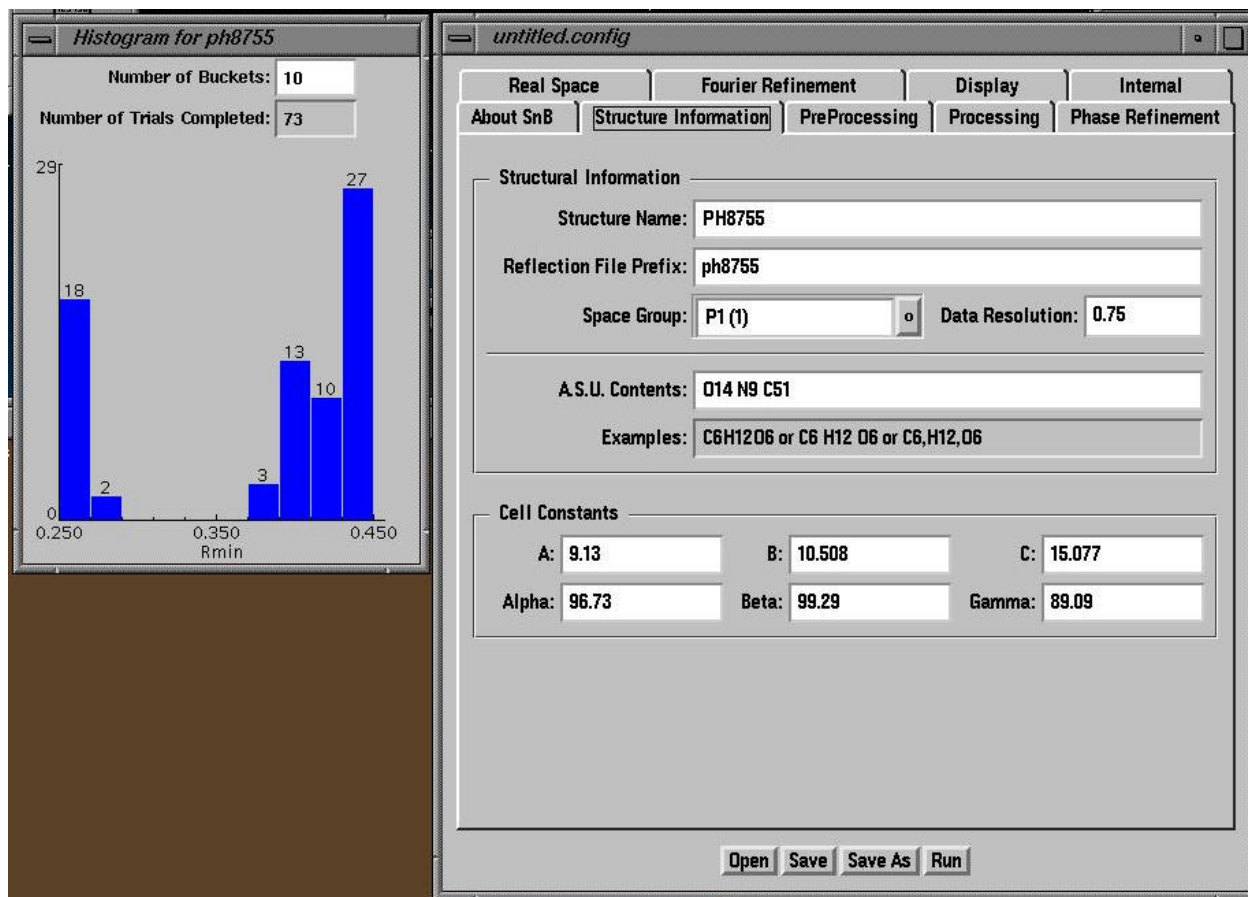


Figure 3. A prototype of the GUI for *SnB* v2.0. The right panel is a snapshot of the main (“Structure Information”) page, while the left panel is a snapshot of the new dynamic histogram tool.

GUI. A prototype of the new *SnB* v2.0 interface is shown on the right frame of Figure 3. The Java language was chosen for this interface due to its extreme portability and ease of management. Once the basic information is typed into the appropriate slots on the “Structure Information” screen, the user is provided with default values for the other necessary parameters. For example, the information given in the right panel of Figure 3 was generated by the system based on extensive experimentation done by the *SnB* research team to determine appropriate values. Of course, the user has the freedom to change any of the default values provided. A modern GUI-based histogram is provided with *SnB* v2.0, as shown on the left of Figure 3. In addition to being graphical, this histogram is dynamic in that it is updated in real time as additional trial structures are processed. The output of the *SnB* v2.0 program has also been made more useful and convenient by the provision of a Java/VRML visualization package [Fass, Miller, & Weeks, 1998], as illustrated in Figure 4. This routine has the benefit of not only allowing the user to view the potential solution as it comes out of *SnB*, but also allowing on-screen editing of the peak/atom file. The revised file can be saved and used as input to either *SnB* or another program for further structure refinement.

Peaks on Special Positions. A second significant deficiency of *SnB* v1.5 was discovered during the investigation of the conotoxin EpI peptide. This structure, which crystallizes in space group

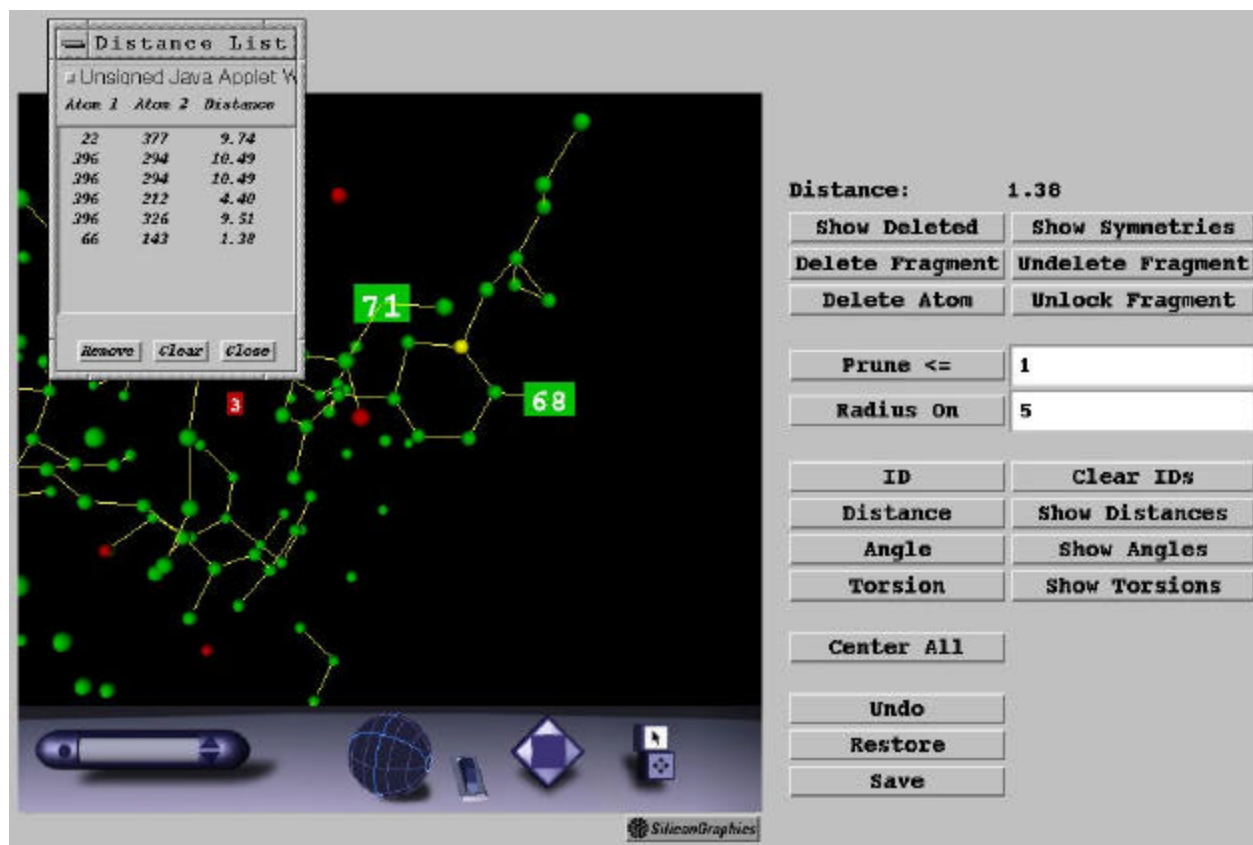


Figure 4. A prototype of the Java/VRML visualization tool provided with *SnB* v2.0.

I4, could not be solved until a patch was put into *SnB* v1.5 that eliminated all peaks within 1.5\AA of any rotation axis. The Se substructure of AdoHcy hydrolase (space group $C222$) was similarly unsolvable until peaks near special positions were eliminated. In addition, once the appropriate patch was in place, the success rate (percentage of trial structures going to solution) for tetragonal vancomycin increased dramatically. It is interesting to note that none of these structures actually has a protein atom located near a special position. The effect of including incorrect peaks at special positions in *SnB* v1.5 is magnified by the fact that there is no provision for assigning proper weights based on multiplicity during the structure-factor calculations. These problems are addressed in *SnB* v2.0 in a manner valid for all space groups by the addition of two new parameters. These parameters are *i*) a minimum distance between symmetry-related peaks such that peaks violating this restriction are eliminated, and *ii*) a maximum number of the highest peaks permitted as exceptions to *i*). The first parameter has a default value of 3.0\AA , and no exceptions are permitted unless some atoms are expected to be on special positions. In situations where such atoms are permitted, they are weighted properly.

Default Parameters. The recommendations for parameter settings will continue to be updated on the Web site as new information surfaces. As of the writing of this document, the default values for many of the critical *SnB* v2.0 parameters are given below. Note that solutions are obtainable for all single diffraction data situations listed when the data resolution is 1.2\AA or higher, especially when atoms like S or Cl are present.

☞ General Parameter Settings

- Given n nonhydrogen atoms (excluding solvent atoms)
- Phase Refinement Method: Parameter Shift
 - Noncentrosymmetric Space Groups
 - ◆ Phase Shift per Phase: 90°
 - ◆ Maximum Number of Attempted Phase Shifts per Phase: 2
 - ◆ Complete Passes Through Set of Reflections: 3 or 1 (P1)
 - Centrosymmetric Space Groups
 - ◆ Phase Shift per Phase: 180°
 - ◆ Maximum Number of Attempted Phase Shifts per Phase: 1
 - ◆ Complete Passes Through Set of Reflections: 1
- Random Seed: Use a 5 digit prime number
- Trials to Generate: Generate 1000 since the time/space is minimal
- Atoms per starting trial: $\text{Min}(n, 100)$
- Fourier Grid Size: $(\text{Resolution of Data})/3$
- Number of E-Fourier cycles: on the order of $0.05n - 0.1n$

☞ Using S_nB at 1.0\AA or Higher

- Phases: $10n$
- Triples:
 - $100n$ if significant high resolution data is available
- Cycles:
 - $n/2$ if $n < 400$ and atoms heavier than C, N, and O are present
 - n otherwise
- Peaks Recycled:
 - $0.4n$ if atoms heavier than C, N, and O are present
 - $0.8n$ otherwise
- Phase Refinement: Parameter Shift

☞ S_nB at $1.1 - 1.4\text{\AA}$

- Increase the number of invariants ($200n - 500n$) and/or
- Perform more cycles (n) of *Shake-and-Bake*
- “Heavy” atoms (*e.g.*, Cl or S) increases the probability of success

☞ S_nB for SAS or SIR Substructure

- DREAR parameters [Smith, Nagar, Rini, Hauptman, & Blessing, 1998]
 - Minimum $F/\sigma(F)$: 3.0
 - Minimum $E/\sigma(E)$: 3.0
 - Minimum $\Delta E/\sigma(\Delta E)$: 1.0
 - Minimum $\text{Diff}E/\sigma(\text{Diff}E)$: 3.0
- Given n substructure atoms
 - Phases: $20n$
 - Invariants: $200n$
 - Cycles: n
 - Peaks recycled: n
 - Phase Refinement: Parameter Shift

Warnings for Substructure Applications. Experience has shown that successful substructure applications are highly dependent on the accuracy of the isomorphous and anomalous normalized difference magnitudes (difference $|E|s$). The amount of data available for these problems is much larger than for full structure problems with a comparable number of atoms to be located. Consequently, the user can afford to be stringent in eliminating data with uncertain measurements. It is important that the suggested guidelines for rejection of such data during processing be met or exceeded. The probability of very large difference $|E|s$ (*e.g.*, > 4) is remote, and data sets that appear to have many such measurements should be examined critically for measurement errors. If a few such data remain even after the adoption of rigorous rejection criteria, it may be best to eliminate them individually.

Conversely, it is also important that a high phase:invariant ratio be maintained in order to insure that the phases are overdetermined. Since the largest $|E|s$ for the substructure cell are more widely separated than they are in a true small-molecule cell, the relative number of possible Σ_2 interactions among the largest reciprocal-lattice vectors can be much smaller. Consequently, a relatively small number of substructure phases (*i.e.*, $10n$) may not have a sufficient number (*i.e.*, $100n$) of invariants. Since the number of interactions increases exponentially with the number of reflections considered, the appropriate action in such cases is to increase the number of reflections to $20n$ (or more). This will typically produce the desired overdetermination. If, however, doing this causes the minimum difference $|E|$ utilized to be too small to be very reliable (*e.g.*, < 1.2), then too many reflections might have been rejected during data processing. In this situation, measurement of more reliable data may be necessary.

V. ACKNOWLEDGMENTS

The authors would like to thank Jimmy Xu, Thomas Tang, Jan Pevzner, and Adam Fass for their contributions to the program, and Herb Hauptman for his continued inspiration and support. This research is supported by grants GM-46733 (NIH) and IRI-9412415 (NSF).

References

- Anderson, D. H., Weiss, M. S., & Eisenberg, D. (1996). A challenging case for protein crystal structure determination: The mating pheromone Er-1 from *Euplotes rakovi*. *Acta Crystallographica* **D52**, 469-480.
- Baggio, R., Woolfson, M. M., Declercq, J.-P., & Germain, G. (1978). On the application of phase relationships to complex structures. XVI. A random approach to structure determination. *Acta Crystallographica* **A34**, 883-892.
- Bhuiya, A. K., & Stanley, E. (1963). The refinement of atomic parameters by direct calculation of the minimum residual. *Acta Crystallographica* (16), 981-984.
- Blessing, R. H., Guo, D. Y., & Langs, D. A. (1996). Statistical expectation value of the Debye-Waller factor and $E(hkl)$ values for macromolecular crystals. *Acta Crystallographica* **D52**, 257-266.
- Chang, C.-S., Weeks, C. M., Miller, R., & Hauptman, H. A. (1997). Incorporating tangent refinement in the Shake-and-Bake formalism. *Acta Crystallographica* **A53**, 436-444.
- Cochran, W. (1955). Relations between the phases of structure factors. *Acta Crystallographica* **8**, 473-478.
- Deacon, A., Weeks, C. M., Miller, R., & Ealick, S. E. (1998). The Shake-and-Bake structure determination of triclinic lysozyme. *Proceedings of the National Academy of Sciences, U.S.A.*, in press.
- Debaerdemaeker, T., & Woolfson, M. M. (1983). On the application of phase relationships to complex structures. XXII. Techniques for random phase refinement. *Acta Crystallographica* **A39**, 193-196.
- DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R., & Hauptman, H. A. (1994). Structure solution by minimal function phase refinement and Fourier filtering. I. Theoretical basis. *Acta Crystallographica* **A50**, 203-210.
- Doublet, S., & Carter, C. W. (1992). *Preparation of selenomethionyl Protein Crystals: Crystallization of Nucleic Acids and Proteins - A Practical Approach*: IRL Press.
- Fass, A., Miller, R., & Weeks, C. M. (1998). *The design and implementation of SnB v2.0 for solving molecular crystal structures*.
- Gallo, S. M., Miller, R., & Weeks, C. M. (1996). *The design of a portable scientific tool: A case study using SnB*. Supercomputing '96 Conference Proceedings.
- Germain, G., Main, P., & Woolfson, M. M. (1971). The application of phase relationships to complex structures. III. The optimum use of phase relationships. *Acta Crystallographica* **A27**, 368-376.
- Germain, G., & Woolfson, M. M. (1968). On the application of phase relationships to complex structures. *Acta Crystallographica* **B24**, 91-96.
- Hauptman, H. A. (1991). A Minimal Principal in the Phase Problem. In D. Moras, A. D. Podjarny, & J. C. Thierry (Eds.), *Crystallographic Computing 5: From Chemistry to Biology* (pp. 324-332): IUCr Oxford University Press.
- Hauptman, H. A. (1995). Looking ahead. *Acta Crystallographica* **B51**, 416-422.
- Hendrickson, W. A., Pahler, A., Smith, J. L., Saton, Y., Merritt, E. A., & Phizackerley, R. P. (1989). Crystal structure of core streptavidin determined from multiwavelength anomalous diffraction of synchrotron radiation. *Proceedings of the National Academy of Sciences, U.S.A.* **86**, 2190-2194.

- Hu, S.-H., Loughnan, M., Miller, R., Weeks, C. M., Blessing, R. H., Alewood, P. F., Lewis, R. J., & Martin, J. L. (1998). The 1.1Å crystal structure of [Tyr¹⁵]-EpI, a novel α -conotoxin from *Conus Episcopatus*, solved by direct methods. *Biochemistry*, in press.
- Hull, S. E., & Irwin, M. J. (1978). On the application of phase relationships to complex structures. XIV. The additional use of statistical information in tangent-formula refinement. *Acta Crystallographica* **A34**, 863-870.
- Karle, J., & Hauptman, H. A. (1956). A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P₁2, 3P₂2. *Acta Crystallographica* **9**, 635-651.
- Loll, P. J., Bevivino, A. E., Korty, B. D., & Axelsen, P. H. (1997). Simultaneous recognition of a carboxylate-containing ligand and intermolecular surrogate ligand in the crystal structure of an asymmetric vancomycin dimer. *Journal of the American Chemical Society* **119**, 1516-1522.
- Loll, P. J., Miller, R., Weeks, C. M., & Axelsen, P. H. (1998). A ligand-mediated dimerization mode for vancomycin. *Chemistry and Biology* **5**, in press.
- Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M., & Hauptman, H. A. (1993). On the application of the minimal principle to solve unknown structures. *Science* **259**, 1430-1433.
- Miller, R., Gallo, S. M., Khalak, H. G., & Weeks, C. M. (1994). SnB: Crystal structure determination via Shake-and-Bake. *Journal of Applied Crystallography* **27**, 613-621.
- Miller, R., & Weeks, C. M. (1997). Shake-and-Bake: Applications and advances. In S. Fortier (Ed.), *Direct Methods for Solving Macromolecular Structures* (pp. 389-400): Kluwer Academic Press.
- Mukherjee, A. K., Helliwell, J. R., & Main, P. (1989). The use of MULTAN to locate the positions of anomalous scatterers. *Acta Crystallographica* **A45**, 715-718.
- Prive, G., Ogihara, N., Wesson, L., Cascio, D., & Eisenberg, E. (1995). *A designer peptide at high resolution: Shake-and-Bake solution of a 400 atom structure*. Proceedings of the American Crystallographic Association Annual Meeting, Montreal, Canada.
- Sheldrick, G. M. (1985). SHELX-84. In G. M. Sheldrick, C. Kruger, & R. Goddard (Eds.), *Crystallographic Computing 3: Data Collection, Structure Determination, Proteins, and Databases* (pp. 184-189). Oxford: Clarendon Press.
- Sheldrick, G. M. (1997). *The SHELX Homepage*. Available: linux.uni-ac.gwdg.de/SHELX/.
- Sheldrick, G. M., & Gould, R. O. (1995). Structure solution by iterative peaklist optimization and tangent expansion in space group P1. *Acta Crystallographica* **B51**, 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). The ab initio structure and refinement of a scorpion protein toxin. *Acta Crystallographica* **D53**, 551-557.
- Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A., & Blessing, R. H. (1998). The use of SnB to determine an anomalous scattering substructure. *Acta Crystallographica*, in press.
- Turner, M. A., Yuan, C.-S., Borchardt, R. T., Hershfield, M. S., Smith, G. D., & Howell, P. L. (1998). Structure determination of selenomethionyl S-adenosylhomocysteine hydrolase using data at a single wavelength. *Nature Structural Biology* **5**, 369-376.
- Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P., & Miller, R. (1994a). Structure solution by minimal function phase refinement and Fourier filtering. II. Implementation and applications. *Acta Crystallographica* **A50**, 210-220.
- Weeks, C. M., Hauptman, H. A., Chang, C.-S., & Miller, R. (1994b). Structure Determination by Shake-and-Bake with Tangent Refinement. In G. Bricogne & C. W. Carter (Eds.), *Likelihood*,

- Bayesian, Interence and Their Application to the Solution of New Structures* (Vol. 30, pp. 153-161): American Crystallographic Association.
- Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M., & Miller, R. (1995). Crambin: A direct solution for a 400 atom structure. *Acta Crystallographica* **D51**, 33-38.
- Weeks, C. M., & Miller, R. (1996). SnB: Applying Shake-and-Bake to Proteins. In P. Bourne & K. Watenpaugh (Eds.), *Crystallographic Computing 7: Proceedings of the Macromolecular Crystallographic Computing School* (pp. 138-147). Bellingham, WA: International Union of Crystallography.
- Weeks, C. M., & Miller, R. (1997). *Macromolecular Phasing by Shake-and-Bake*. Recent Advances in Phasing: Proceedings of the CCP4 Study Weekend (DL-CONF-97-001), Warrington, UK.
- Weeks, C. M., & Miller, R. (1998). The design and implementation of SnB v2.0. *Journal of Applied Crystallography*, submitted.
- Wilson, K. S. (1978). The application of MULTAN to the analysis of isomorphous derivatives in protein crystallography. *Acta Crystallographica* **B34**, 1599-1608.
- Yao, J.-X. (1981). On the application of phase relationships to complex structures. XVIII. RANTAN - random MULTAN. *Acta Crystallographica* **A37**, 642-644.

SnB v2.0 Basic Data: simple structure

GENERAL INFORMATION

Title:

Space Group: • Data Type: •

Asymmetric Unit Contents (SAS or SIR Substructure):

Cell Constants
 A: B: C:
 Alpha: Beta: Gamma:

Radiation: • Wavelength: •

Number of anomalous dispersion correction terms:

Element: f: f':

CREATE E's (DREAR Interface)

Native Input File : File Type: •

Derivative Input File : File Type: •

Output File Name (for input to *SnB*):

ASU Contents: Native Derivative

Use Bayesian E's? Yes No Minimum F/σ(F) for local scaling:

DiffE Limits: Data Resolution Minimum Maximum

Minimums: E/σ(E) ΔE/σ(ΔE) DiffE/σ(DiffE)

Execute DREAR Suite

View DREAR Results

Clean DREAR Files

View DREAR Documentation

REFL & INVInput Reflection File Name: ◆ New to *SnB* ◇ Old *SnB* FileInput Invariant File Name: ◆ New ◇ ExistingDate Resolution: Minimum Maximum Number of Reflections to use: Number of Triples to Use: **TRIALS & CYCLES**

Starting Phases from: ◆ Random Atoms ◇ Random Phases

◇ Variable Input Phases ◇ Fixed Input Phases ◇ Model Structure Atoms

Number of Trials: Starting at Trial: Random Seed (Prime): •Input Phase File Name: Input Atom File Name: Number of *Shake-and-Bake* Cycles: Terminate trials failing the R-Ratio test? ◇ Yes ◆ No R-Ratio Cutoff: **PHASE REFINEMENT**

Method: ◆ Parameter Shift ◇ Tangent Formula

Parameter Shift Options: Phase Shift: Number of Shifts: Number of passes through phase set: Tangent Formula Options: Number of passes through phase set:

CONSTRAINTSNumber Of Peaks To Select: Fourier Grid Size (map resolution): Minimum Interpeak Distance: Minimum distance between symmetry-related peaks (defines special position excluded volume): Number of special position peaks to keep: Perform extra cycles with more peaks? Yes NoNumber of Extra Cycles: Number of Peaks: **TWICE BAKING**Trials For E-Fourier Filtering (Fourier Refinement)? None All Best OnlyNumber of Cycles: Number of Peaks: F/ σ (F) Cutoff: Minimum |E|: **PROCESS TRIALS**File name prefix for results: Keep complete trace file ? Yes No Number of *SnB* jobs to submit :
(every cycle) (processors available)**DISPLAY**Result files prefix: Number of peaks to use: Maximum bond distance (\AA): Number of large peaks to be distinguished: Maximum bond distance (\AA) for large peaks:

SnB* v2.0 Basic Data: difficult structure job 1*GENERAL INFORMATION**Title: Space Group: • Data Type: •Asymmetric Unit Contents (SAS or SIR Substructure): Cell Constants
A: B: C:
Alpha: Beta: Gamma: Radiation: • Wavelength: •Number of anomalous dispersion correction terms: Element: f: f': **CREATE E's (DREAR Interface)**Native Input File: File Type: •Derivative Input File: File Type: •Output File Name (for input to *SnB*): ASU Contents: Native Derivative Use Bayesian E's? Yes No Minimum F/σ(F) for local scaling: DiffE Limits: Data Resolution Minimum Maximum Minimums: E/σ(E) ΔE/σ(ΔE) DiffE/σ(DiffE)

REFL & INVInput Reflection File Name: ◆ New to *SnB* ◇ Old *SnB* FileInput Invariant File Name: ◆ New ◇ ExistingDate Resolution: Minimum Maximum Number of Reflections to use: Number of Triples to Use: **TRIALS & CYCLES**

Starting Phases from: ◆ Random Atoms ◇ Random Phases

◇ Variable Input Phases ◇ Fixed Input Phases ◇ Model Structure Atoms

Number of Trials: Starting at Trial: Random Seed (Prime): •Input Phase File Name: Input Atom File Name: Number of *Shake-and-Bake* Cycles: Terminate trials failing the R-Ratio test? ◇ Yes ◆ No R-Ratio Cutoff: **PHASE REFINEMENT**

Method: ◆ Parameter Shift ◇ Tangent Formula

Parameter Shift Options: Phase Shift: Number of Shifts: Number of passes through phase set: Tangent Formula Options: Number of passes through phase set:

CONSTRAINTSNumber Of Peaks To Select: Fourier Grid Size (map resolution): Minimum Interpeak Distance: Minimum distance between symmetry-related peaks (defines special position excluded volume): Number of special position peaks to keep: Perform extra cycles with more peaks? Yes NoNumber of Extra Cycles: Number of Peaks: **TWICE BAKING**Trials For E-Fourier Filtering (Fourier Refinement)? None All Best OnlyNumber of Cycles: Number of Peaks: F/ σ (F) Cutoff: Minimum |E|: **PROCESS TRIALS**File name prefix for results: Keep complete trace file ? Yes No Number of *SnB* jobs to submit :
(every cycle) (processors available)**DISPLAY**Result files prefix: Number of peaks to use: Maximum bond distance (\AA): Number of large peaks to be distinguished: Maximum bond distance (\AA) for large peaks:

SnB v2.0 Basic Data: difficult structure job 2

(Use old reflection and invariant files. Use a single 112.5° phase shift.
Do more trials, with early termination for those failing the R-Ratio test.)

REFL & INV

Input Reflection File Name: New to SnB Old SnB File

Input Invariant File Name: New Existing

Date Resolution: Minimum Maximum

Number of Reflections to use: Number of Triples to Use:

TRIALS & CYCLES

Starting Phases from: Random Atoms Random Phases
 Variable Input Phases Fixed Input Phases Model Structure Atoms

Number of Trials: Starting at Trial: Random Seed (Prime):

Input Phase File Name:

Input Atom File Name:

Number of *Shake-and-Bake* Cycles:

Terminate trials failing the R-Ratio test? Yes No R-Ratio Cutoff:

PHASE REFINEMENT

Method: Parameter Shift Tangent Formula

Parameter Shift Options: Phase Shift: Number of Shifts:

Number of passes through phase set:

Tangent Formula Options: Number of passes through phase set:

PROCESS TRIALS

File name prefix for results:

SnB v2.0 Basic Data: difficult structure job 3(Repeat the best trial from job 2: perform extra *Shake-and-Bake* cycles and Fourier refinement.)**TRIALS & CYCLES**

Starting Phases from:

 Random Atoms Random Phases Variable Input Phases Fixed Input Phases Model Structure AtomsNumber of Trials: Starting at Trial: Random Seed (Prime): Input Phase File Name: Input Atom File Name: Number of *Shake-and-Bake* Cycles: Terminate trials failing the R-Ratio test? Yes No R-Ratio Cutoff: **CONSTRAINTS**Number Of Peaks To Select: Fourier Grid Size (map resolution): Minimum Interpeak Distance: Minimum distance between symmetry-related peaks (defines special position excluded volume): Number of special position peaks to keep: Perform extra cycles with more peaks? Yes NoNumber of Extra Cycles: Number of Peaks: **TWICE BAKING**Trials For E-Fourier Filtering (Fourier Refinement)? None All Best OnlyNumber of Cycles: Number of Peaks: F/ σ (F) Cutoff: Minimum |E|: **PROCESS TRIALS**File name prefix for results:

SnB v2.0 Basic Data: difficult structure job 4

(Use the VRML visualizer to edit the atom file. Do additional Fourier refinement to improve the model structure which starts with 250 atoms.)

TRIALS & CYCLES

Starting Phases from:

 Random Atoms Random Phases Variable Input Phases Fixed Input Phases Model Structure Atoms

Number of Trials:

Starting at Trial:

Random Seed (Prime):

Input Phase File Name:

Input Atom File Name:

Number of *Shake-and-Bake* Cycles:Terminate trials failing the R-Ratio test? Yes No

R-Ratio Cutoff:

CONSTRAINTS

Number Of Peaks To Select:

Fourier Grid Size (map resolution):

Minimum Interpeak Distance:

Minimum distance between symmetry-related peaks (defines special position excluded volume):

Number of special position peaks to keep:

Perform extra cycles with more peaks?

 Yes No

Number of Extra Cycles:

Number of Peaks:

TWICE BAKING

Trials For E-Fourier Filtering (Fourier Refinement)?

 None All Best Only

Number of Cycles:

Number of Peaks:

F/ σ (F) Cutoff:

Minimum |E|:

PROCESS TRIALS

File name prefix for results:

SnB* v2.0 SAS Data*GENERAL INFORMATION**Title: Space Group: • Data Type: •Asymmetric Unit Contents (SAS or SIR Substructure): Cell Constants
A: B: C:
Alpha: Beta: Gamma: Radiation: • Wavelength: •Number of anomalous dispersion correction terms: Element: f: f': **CREATE E's (DREAR Interface)**Native Input File : File Type: •Derivative Input File : File Type: •Output File Name (for input to *SnB*): ASU Contents: Native Derivative Use Bayesian E's? Yes No Minimum F/σ(F) for local scaling: DiffE Limits: Data Resolution Minimum Maximum Minimums: E/σ(E) ΔE/σ(ΔE) DiffE/σ(DiffE)

REFL & INVInput Reflection File Name: ◆ New to *SnB* ◇ Old *SnB* FileInput Invariant File Name: ◆ New ◇ ExistingDate Resolution: Minimum Maximum Number of Reflections to use: Number of Triples to Use: **TRIALS & CYCLES**

Starting Phases from: ◆ Random Atoms ◇ Random Phases

◇ Variable Input Phases ◇ Fixed Input Phases ◇ Model Structure Atoms

Number of Trials: Starting at Trial: Random Seed (Prime): •Input Phase File Name: Input Atom File Name: Number of *Shake-and-Bake* Cycles: Terminate trials failing the R-Ratio test? ◇ Yes ◆ No R-Ratio Cutoff: **PHASE REFINEMENT**

Method: ◆ Parameter Shift ◇ Tangent Formula

Parameter Shift Options: Phase Shift: Number of Shifts: Number of passes through phase set: Tangent Formula Options: Number of passes through phase set:

CONSTRAINTSNumber Of Peaks To Select: Fourier Grid Size (map resolution): Minimum Interpeak Distance: Minimum distance between symmetry-related peaks (defines special position excluded volume): Number of special position peaks to keep: Perform extra cycles with more peaks? Yes NoNumber of Extra Cycles: Number of Peaks: **TWICE BAKING**Trials For E-Fourier Filtering (Fourier Refinement)? None All Best OnlyNumber of Cycles: Number of Peaks: F/ σ (F) Cutoff: Minimum |E|: **PROCESS TRIALS**File name prefix for results: Keep complete trace file ? Yes No Number of *SnB* jobs to submit :
(every cycle) (processors available)**DISPLAY**Result files prefix: Number of peaks to use: Maximum bond distance (\AA): Number of large peaks to be distinguished: Maximum bond distance (\AA) for large peaks:

SnB v2.0 SAS Data: job 2

(Atom:phase:invariant ratio changed to 1:20:200 because of failure to generate the requested number of invariants with ratio 1:10:100.)

REFL & INV

Input Reflection File Name: New to *SnB* Old *SnB* File

Input Invariant File Name: New Existing

Date Resolution: Minimum Maximum

Number of Reflections to use: Number of Triples to Use:

PROCESS TRIALS

File name prefix for results:

Keep complete trace file ? Yes No Number of *SnB* jobs to submit :
(every cycle) (processors available)

SnB* v2.0 SIR Data: job 1*GENERAL INFORMATION**Title: Space Group: • Data Type: •Asymmetric Unit Contents (SAS or SIR Substructure): Cell Constants
A: B: C:
Alpha: Beta: Gamma: Radiation: • Wavelength: •Number of anomalous dispersion correction terms: Element: f: f': **CREATE E's (DREAR Interface)**Native Input File : File Type: •Derivative Input File : File Type: •Output File Name (for input to *SnB*): ASU Contents: Native Derivative Use Bayesian E's? Yes No Minimum F/σ(F) for local scaling: DiffE Limits: Data Resolution Minimum Maximum Minimums: E/σ(E) ΔE/σ(ΔE) DiffE/σ(DiffE)

REFL & INVInput Reflection File Name: ◆ New to *SnB* ◇ Old *SnB* FileInput Invariant File Name: ◆ New ◇ ExistingDate Resolution: Minimum Maximum Number of Reflections to use: Number of Triples to Use: **TRIALS & CYCLES**

Starting Phases from: ◆ Random Atoms ◇ Random Phases

◇ Variable Input Phases ◇ Fixed Input Phases ◇ Model Structure Atoms

Number of Trials: Starting at Trial: Random Seed (Prime): •Input Phase File Name: Input Atom File Name: Number of *Shake-and-Bake* Cycles: Terminate trials failing the R-Ratio test? ◇ Yes ◆ No R-Ratio Cutoff: **PHASE REFINEMENT**

Method: ◆ Parameter Shift ◇ Tangent Formula

Parameter Shift Options: Phase Shift: Number of Shifts: Number of passes through phase set: Tangent Formula Options: Number of passes through phase set:

CONSTRAINTSNumber Of Peaks To Select: Fourier Grid Size (map resolution): Minimum Interpeak Distance: Minimum distance between symmetry-related peaks (defines special position excluded volume): Number of special position peaks to keep: Perform extra cycles with more peaks? Yes NoNumber of Extra Cycles: Number of Peaks: **TWICE BAKING**Trials For E-Fourier Filtering (Fourier Refinement)? None All Best OnlyNumber of Cycles: Number of Peaks: F/ σ (F) Cutoff: Minimum |E|: **PROCESS TRIALS**File name prefix for results: Keep complete trace file ? Yes No Number of *SnB* jobs to submit :
(every cycle) (processors available)**DISPLAY**Result files prefix: Number of peaks to use: Maximum bond distance (\AA): Number of large peaks to be distinguished: Maximum bond distance (\AA) for large peaks:

