An Introduction to SnB v2.0

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

\textsuperscript{a}The Hauptman-Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203
\textsuperscript{b}Department of Computer Science and Engineering, SUNY-Buffalo, Buffalo NY 14260
\{miller,weeks\}@hwi.buffalo.edu; www.hwi.buffalo.edu/SnB/

Abstract. SnB is a computer program based on \textit{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, \textit{SnB} has also been used to determine the Se sites in large selenomethionyl-substituted proteins. \textit{SnB} v1.5 has been available for several years and is being used regularly in many laboratories. At this workshop, we introduce \textit{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, \textit{SnB} v2.0 provides the user with a variety of new algorithmic options.

I. INTRODUCTION

\textit{SnB} is a publicly available direct-methods package based on the \textit{Shake-and-Bake} method of structure determination. The program has been available since 1994 and has been available for download from the \textit{SnB} Web site\textsuperscript{1} 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, \textit{SnB} represented a significant advance in \textit{ab initio} direct-methods phasing. In fact, due to the success of \textit{SnB}, Sheldrick has recently exploited the \textit{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, \textit{SnB} has also been used to increase the number of Se sites that can be located for selenomethionyl-substituted proteins. For example, \textit{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 \textit{Shake-and-Bake} procedure. In Section III, we give an overview of the \textit{SnB} program, including features available in the current public release of the program, \textit{SnB} v1.5, and details of implementation. We also discuss new features that are incorporated into \textit{SnB} v2.0, including a graphical user interface (GUI), a graphical histogram display, an interactive visualization routine, optimizations to the \textit{Shake-and-Bake} procedure, and an interface to the DREAR suite of data-processing programs. An appendix is included that contains examples of using \textit{SnB} v2.0 for traditional single data sets, SAS, and SIR situations.

\textsuperscript{1} www.hwi.buffalo.edu/SnB/
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

\[ \tan(f_n) = \frac{\sum_k |E_k E_{n-k}| \sin(f_k + f_{n-k})}{\sum_k |E_k E_{n-k}| \cos(f_k + f_{n-k})} \]  

[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

![Figure 1](image-url). 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.
[Bhuiya & Stanley, 1963] to reduce the value of an objective function such as the minimal function

\[
R(f) = \frac{\sum_{h,k} A_{hk} \left\{ \cos(f_h + f_k + f_{-h-k}) - \frac{I_1(A_{hk})}{I_0(A_{hk})} \right\}^2}{\sum_{h,k} A_{hk}}
\]

(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)

\[
A_{hk} = \frac{2|E_h E_k E_{-h-k}|}{N^{1/2}}
\]

(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
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) [Doublie & 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

### Table 1

<table>
<thead>
<tr>
<th>PROTEIN STRUCTURE</th>
<th>LOCATION</th>
<th>ATOMS</th>
<th>SPACE GROUP</th>
<th>SUCCESS RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Vancomycin (Tetragonal)</td>
<td>U. Penn</td>
<td>258</td>
<td>P4_3212</td>
<td>0.8%</td>
</tr>
<tr>
<td>*Conotoxin EpI</td>
<td>HWI</td>
<td>289</td>
<td>I4</td>
<td>53.0</td>
</tr>
<tr>
<td>Gramicidin A</td>
<td>HWI</td>
<td>317</td>
<td>P2_12_1</td>
<td>1.1</td>
</tr>
<tr>
<td>*Er-1 pheromone</td>
<td>UCLA</td>
<td>328</td>
<td>C2</td>
<td>0.25</td>
</tr>
<tr>
<td>Crambin</td>
<td>HWI</td>
<td>~400</td>
<td>P2_1</td>
<td>4.8</td>
</tr>
<tr>
<td>*Alpha-1 peptide</td>
<td>OCI/Toronto</td>
<td>471</td>
<td>P1</td>
<td>5.0</td>
</tr>
<tr>
<td>Rubredoxin</td>
<td>HWI</td>
<td>497</td>
<td>P2_1</td>
<td>6.2</td>
</tr>
<tr>
<td>*Vancomycin (Triclinic)</td>
<td>HWI/U. Penn</td>
<td>547</td>
<td>P1</td>
<td>N.A.</td>
</tr>
<tr>
<td>Scorpion Toxin II</td>
<td>HWI</td>
<td>624</td>
<td>P2_12_1</td>
<td>1.4</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>Cornell/HWI</td>
<td>~1200</td>
<td>P1</td>
<td>22.0</td>
</tr>
</tbody>
</table>

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.

### Table 2

<table>
<thead>
<tr>
<th>STRUCTURE</th>
<th>LOCATION</th>
<th>SE ATOMS</th>
<th>PROTEIN SIZE (ASU)</th>
<th>SPACE GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3d</td>
<td>Toronto/HWI</td>
<td>8</td>
<td>34kDA</td>
<td>P2_12_1</td>
</tr>
<tr>
<td>GPATase</td>
<td>Purdue/HWI</td>
<td>22</td>
<td>112</td>
<td>C222_1</td>
</tr>
<tr>
<td>*AdoHcy</td>
<td>Toronto/HWI</td>
<td>30</td>
<td>95</td>
<td>C222</td>
</tr>
</tbody>
</table>

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.
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 \textit{SnB} with \textit{i}) basic information about the crystal (\textit{e.g.}, its chemical contents) and \textit{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 \textit{SnB} Web site. The relative efficiency of tangent-formula and parameter-shift phase refinement in \textit{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 \textit{Shake-and-Bake} phasing procedure is initiated, as follows.
   a) The phases are refined \textit{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 \textit{via} peak selection, and structure-factor calculation is repeated for the predetermined number of \textit{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 graphical visualization tool to aid in the diagnosis of solution,
5. 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.
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 Epl peptide. This structure, which crystallizes in space group
I4, could not be solved until a patch was put into SnB v1.5 that eliminated all peaks within 1.5 Å 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 Å, 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 Å 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: (Resolution of Data)/3
- Number of E-Fourier cycles: on the order of 0.05 \( n \) - 0.1 \( n \)

**Using SnB at 1.0 Å or Higher**

- Phases: 10 \( n \)
- Triples:
  - 100\( n \) 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:
  - 0.4\( n \) if atoms heavier than C, N, and O are present
  - 0.8\( n \) otherwise
- Phase Refinement: Parameter Shift

**SnB at 1.1 - 1.4Å**

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

**SnB 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 \( \text{DiffE}/\sigma(\text{DiffE}) \): 3.0
- Given \( n \) substructure atoms
  - Phases: 20\( n \)
  - Invariants: 200\( n \)
  - 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 $\Sigma_2$ 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


Fass, A., Miller, R., & Weeks, C. M. (1998). *The design and implementation of SnB v2.0 for solving molecular crystal structures*.


SnB v2.0 Basic Data: simple structure

GENERAL INFORMATION

Title: Isoleucinomycin

Space Group: P212121(19) Data Type: Basic

Asymmetric Unit Contents (SAS or SIR Substructure): C60 H102 N6 O18

Cell Constants
A: 11.516  B: 15.705  C: 39.310
Alpha: 90.00  Beta: 90.00  Gamma: 90.00

Radiation: CU Wavelength: 1.5418

Number of anomalous dispersion correction terms:

Element: f': f'':

CREATE E's (DREAR Interface)

Native Input File: iled.hkl File Type: IH,IK,IL,F,σ(F)

Derivative Input File: File Type:

Output File Name (for input to SnB): iled.ref.orig

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 & INV

Input Reflection File Name: iled.ref.orig  ♦ New to SnB  ◊ Old SnB File

Input Invariant File Name:  ❯ New  ◊ Existing

Date Resolution:          Minimum 999.00       Maximum 0.75

Number of Reflections to use: 84 Number of Triples to Use: 8400

TRIALS & CYCLES

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

Number of Trials: 1000    Starting at Trial: 1    Random Seed (Prime): 11909

Input Phase File Name:               

Input Atom File Name:               

Number of Shake-and-Bake Cycles: 42

Terminate trials failing the R-Ratio test? ◊ Yes  ❯ No  R-Ratio Cutoff: 0.23

PHASE REFINEMENT

Method:        ♦ Parameter Shift  ◊ Tangent Formula

Parameter Shift Options:  Phase Shift: 90.0    Number of Shifts: 2

                                   Number of passes through phase set: 3

Tangent Formula Options:  Number of passes through phase set: 

16
CONSTRANTS

Number Of Peaks To Select: 84

Fourier Grid Size (map resolution): 0.33 Minimum Interpeak Distance: 1.00

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: 4 Number of Peaks: 84

F/σ(F) Cutoff: 4.00 Minimum |E|: 0.75

PROCESS TRIALS

File name prefix for results: job1

Keep complete trace file? ◊ Yes ◆ No (every cycle) Number of SnB jobs to submit: 1

(DISPLAY)

Result files prefix: job1

Number of peaks to use: 100 Maximum bond distance (Å): 1.80

Number of large peaks to be distinguished: 0 Maximum bond distance (Å) for large peaks:

DISPLAY HISTOGRAM VISUALIZE STRUCTURE

VIEW GEOMETRY LISTING
**SnB v2.0 Basic Data: difficult structure job 1**

**GENERAL INFORMATION**

Title: Triclinic Vancomycin

Space Group: P1(1)  
Data Type: Basic

Asymmetric Unit Contents (SAS or SIR Substructure): C264 H304 N36 O96 CL8

Cell Constants
- A: 21.50
- B: 24.50
- C: 25.00
- Alpha: 64.00
- Beta: 90.00
- Gamma: 81.00

Radiation: CU  
Wavelength: 1.5418

Number of anomalous dispersion correction terms: 

Element: 
- f': 
- f'': 

**CREATE E's (DREAR Interface)**

Native Input File: trivanco.hkl  
File Type: IH,IK,IL,F,σ(F)

Derivative Input File: 
File Type: 

Output File Name (for input to SnB): trivanco.ref.orig

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 & INV**

Input Reflection File Name: trivanco.ref.orig ◆ New to SnB ◊ Old SnB File

Input Invariant File Name: ◆ New ◊ Existing

Date Resolution: Minimum 999.00 Maximum 0.97

Number of Reflections to use: 4000 Number of Triples to Use: 40000

**TRIALS & CYCLES**

Starting Phases from: ◆ Random Atoms ◊ Random Phases

◊ Variable Input Phases ◊ Fixed Input Phases ◊ Model Structure Atoms

Number of Trials: 1000 Starting at Trial: 1 Random Seed (Prime): 11909 ●

Input Phase File Name:

Input Atom File Name:

Number of Shake-and-Bake Cycles: 400

Terminate trials failing the R-Ratio test? ◊ Yes ◆ No R-Ratio Cutoff: 0.23

**PHASE REFINEMENT**

Method: ◆ Parameter Shift ◊ Tangent Formula

Parameter Shift Options: Phase Shift: 90.0 Number of Shifts: 2

Number of passes through phase set: 1

Tangent Formula Options: Number of passes through phase set:
## CONSTRAINTS

<table>
<thead>
<tr>
<th>Number Of Peaks To Select:</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourier Grid Size (map resolution):</td>
<td>0.33</td>
</tr>
<tr>
<td>Minimum Interpeak Distance:</td>
<td>1.00</td>
</tr>
<tr>
<td>Minimum distance between symmetry-related peaks (defines special position excluded volume):</td>
<td></td>
</tr>
<tr>
<td>Number of special position peaks to keep:</td>
<td></td>
</tr>
<tr>
<td>Perform extra cycles with more peaks?</td>
<td>◯ Yes  ◆ No</td>
</tr>
<tr>
<td>Number of Extra Cycles:</td>
<td></td>
</tr>
<tr>
<td>Number of Peaks:</td>
<td></td>
</tr>
</tbody>
</table>

## TWICE BAKING

<table>
<thead>
<tr>
<th>Trials For E-Fourier Filtering (Fourier Refinement)?</th>
<th>◆ None  ◯ All  ◯ Best Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Cycles:</td>
<td></td>
</tr>
<tr>
<td>Number of Peaks:</td>
<td></td>
</tr>
<tr>
<td>F/σ(F) Cutoff:</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
</tr>
</tbody>
</table>

## PROCESS TRIALS

<table>
<thead>
<tr>
<th>File name prefix for results:</th>
<th>job1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keep complete trace file?</td>
<td>◯ Yes  ◆ No</td>
</tr>
<tr>
<td>(every cycle)</td>
<td></td>
</tr>
<tr>
<td>Number of SnB jobs to submit:</td>
<td>1</td>
</tr>
<tr>
<td>(processors available)</td>
<td></td>
</tr>
</tbody>
</table>

## DISPLAY

<table>
<thead>
<tr>
<th>Result files prefix:</th>
<th>job1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of peaks to use:</td>
<td>500</td>
</tr>
<tr>
<td>Maximum bond distance (Å):</td>
<td>1.80</td>
</tr>
<tr>
<td>Number of large peaks to be distinguished:</td>
<td>8</td>
</tr>
<tr>
<td>Maximum bond distance (Å) for large peaks:</td>
<td>2.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DISPLAY HISTOGRAM</th>
<th>VISUALIZE STRUCTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIEW GEOMETRY LISTING</td>
<td></td>
</tr>
</tbody>
</table>
**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: `job1.SnB_ref` ◊ New to SnB ◆ Old SnB File

Input Invariant File Name: `job1.SnB_inv` ◊ New ◆ Existing

Date Resolution: ◆ Minimum 999.00 ◆ Maximum 0.97

Number of Reflections to use: 4000 Number of Triples to Use: 40000

**TRIALS & CYCLES**

Starting Phases from: ◆ Random Atoms ◆ Random Phases

◊ Variable Input Phases ◆ Fixed Input Phases ◆ Model Structure Atoms

Number of Trials: 5000 Starting at Trial: 1 Random Seed (Prime): 11909

Input Phase File Name:

Input Atom File Name:

Number of Shake-and-Bake Cycles: 400

Terminate trials failing the R-Ratio test? ◆ Yes ◆ No R-Ratio Cutoff: 0.23

**PHASE REFINEMENT**

Method: ◆ Parameter Shift ◆ Tangent Formula

Parameter Shift Options: Phase Shift: 112.50 Number of Shifts: 1

Number of passes through phase set: 1

Tangent Formula Options: Number of passes through phase set:

**PROCESS TRIALS**

File name prefix for results: `job2`
**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 Atoms

Number of Trials: 1  Starting at Trial: 3456  Random Seed (Prime): 11909

Input Phase File Name: 
Input Atom File Name: 

Number of Shake-and-Bake Cycles: 400

Terminate trials failing the R-Ratio test? ◆ Yes ◄ No  R-Ratio Cutoff: 0.23

**CONSTRAINTS**

Number Of Peaks To Select: 150

Fourier Grid Size (map resolution): 0.33  Minimum Interpeak Distance: 1.00

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: 40  Number of Peaks: 400

**TWICE BAKING**

Trials For E-Fourier Filtering (Fourier Refinement)? ◆ None ◄ All ◄ Best Only

Number of Cycles: 40  Number of Peaks: 400

F/σ(F) Cutoff: 4.00  Minimum |E|: 0.75

**PROCESS TRIALS**

File name prefix for results: job3
**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: 1 Starting at Trial: 1 Random Seed (Prime):

Input Phase File Name: 

Input Atom File Name: job3.SnB_atom

Number of Shake-and-Bake Cycles: 0

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

**CONSTRAINTS**

Number Of Peaks To Select: 250

Fourier Grid Size (map resolution): 0.33 Minimum Interpeak Distance: 1.00

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: 40 Number of Peaks: 400

**TWICE BAKING**

Trials For E-Fourier Filtering (Fourier Refinement)? ◦ None ◦ All ◦ Best Only

Number of Cycles: 40 Number of Peaks: 400

F/σ(F) Cutoff: 4.0 Minimum |E|: 0.5

**PROCESS TRIALS**

File name prefix for results: job4
**SnB v2.0 SAS Data**

**GENERAL INFORMATION**

**Title:** S-Adenosylhomocysteine (AdoHcy) hydrolase peak data

**Space Group:** C222(21)  **Data Type:** SAS

**Asymmetric Unit Contents (SAS or SIR Substructure):** Se30

**Cell Constants**
- A: 91.93
- B: 168.02
- C: 137.77
- Alpha: 90.0
- Beta: 90.0
- Gamma: 90.0

**Radiation:** SYNCHROTRON  **Wavelength:** 0.9784

**Number of anomalous dispersion correction terms:** 1

**Element:** Se  **f':** -7.35  **f'':** 5.92

**CREATE E's (DREAR Interface)**

**Native Input File:** AdoHcy.hkl  **File Type:** IH,IK,IL,F,σ(F)

**Derivative Input File:**  **File Type:**

**Output File Name (for input to SnB):** AdoHcy.ref.orig

**ASU Contents:** Native C4289 H6767 N1169 O1276 S20 Se30 P4  Derivative

**Use Bayesian E's?** ♦ Yes  ♦ No  **Minimum F/σ(F) for local scaling:** 3.0

**DiffE Limits:**  **Data Resolution Minimum:** 999.00  **Maximum:** 0.75

**Minimums:**  **E/σ(E):** 3.0  **ΔE/σ(ΔE):** 1.0  **DiffE/σ(DiffE):** 3.0

**Execute DREAR Suite**  **View DREAR Results**

**Clean DREAR Files**  **View DREAR Documentation**
**REFL & INV**

Input Reflection File Name: **AdoHcy.ref.orig**  ◆ New to SnB  ◦ Old SnB File

Input Invariant File Name:  ◆ New  ◦ Existing

Date Resolution:  
Minimum 999.00  
Maximum 2.80

Number of Reflections to use: 300  Number of Triples to Use: 3000

**TRIALS & CYCLES**

Starting Phases from:  
◆ Random Atoms  ◦ Random Phases

  ◦ Variable Input Phases  ◦ Fixed Input Phases  ◦ Model Structure Atoms

Number of Trials: 1000  Starting at Trial: 1  Random Seed (Prime): 11909  ◆

Input Phase File Name: 

Input Atom File Name: 

Number of Shake-and-Bake Cycles: 30

Terminate trials failing the R-Ratio test? ◦ Yes  ◆ No  R-Ratio Cutoff: 0.23

**PHASE REFINEMENT**

Method:  
◆ Parameter Shift  ◦ Tangent Formula

Parameter Shift Options:  
Phase Shift: 90.0  Number of Shifts: 2

  Number of passes through phase set: 3

Tangent Formula Options:  
Number of passes through phase set: 

25
CONSTRANTS

Number Of Peaks To Select: 30
Fourier Grid Size (map resolution): 0.93 Minimum Interpeak Distance: 2.80
Minimum distance between symmetry-related peaks (defines special position excluded volume): 3.00 Number of special position peaks to keep: 0
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: job1
Keep complete trace file? ◊ Yes ◆ No (every cycle)
Number of SnB jobs to submit: 1
(processors available)

DISPLAY

Result files prefix: job1
Number of peaks to use: 0 Maximum bond distance (Å):
Number of large peaks to be distinguished: 0 Maximum bond distance (Å) for large peaks:

DISPLAY HISTOGRAM VISUALIZE STRUCTURE

VIEW GEOMETRY LISTING
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: job1.SnB_ref ◊ New to SnB  ◆ Old SnB File
Input Invariant File Name: ◆ New ◊ Existing
Date Resolution: Minimum 999.00 Maximum 2.80
Number of Reflections to use: 600 Number of Triples to Use: 6000

PROCESS TRIALS

File name prefix for results: job2
Keep complete trace file? ◊ Yes ◆ No (every cycle)
Number of SnB jobs to submit: 1 (processors available)

Run
**SnB v2.0 SIR Data: job 1**

**GENERAL INFORMATION**

<table>
<thead>
<tr>
<th>Title:</th>
<th>20B-Hydroxysteroid Dehydrogenase Pt Derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space Group:</td>
<td>P41212(92)</td>
</tr>
<tr>
<td>Data Type:</td>
<td>SIR</td>
</tr>
<tr>
<td>Asymmetric Unit Contents (SAS or SIR Substructure):</td>
<td>Pt4</td>
</tr>
<tr>
<td>Cell Constants</td>
<td>A: 58.53</td>
</tr>
<tr>
<td></td>
<td>Alpha: 90.00</td>
</tr>
<tr>
<td>Radiation:</td>
<td>CU</td>
</tr>
<tr>
<td>Wavelength:</td>
<td>1.5418</td>
</tr>
</tbody>
</table>

Number of anomalous dispersion correction terms: [ ]

<table>
<thead>
<tr>
<th>Element:</th>
<th>f':</th>
<th>f'':</th>
</tr>
</thead>
</table>

**CREATE E's (DREAR Interface)**

<table>
<thead>
<tr>
<th>Native Input File:</th>
<th>20BHSD.nat.hkl</th>
<th>File Type:</th>
<th>IH,IK,IL,F,σ(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derivative Input File:</td>
<td>20BHSD.der.hk</td>
<td>File Type:</td>
<td>IH,IK,IL,F,σ(F)</td>
</tr>
<tr>
<td>Output File Name (for input to SnB):</td>
<td>20BHSD.ref.orig</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| ASU Contents: Native | C1338 H2157 N387 O407 S11 |
| Derivative | C1338 H2157 N387 O407 S11 Pt2 I2 |

| Use Bayesian E's? | Yes | No |
| Minimum F/σ(F) for local scaling: | 3.0 |

| DiffE Limits: | Data Resolution Minimum | 999.00 | Maximum | 0.75 |
| Minimums: | E/σ(E) | 3.0 | ΔE/σ(ΔE) | 1.0 | DiffE/σ(DiffE) | 3.0 |

**Buttons:**

- Execute DREAR Suite
- View DREAR Results
- Clean DREAR Files
- View DREAR Documentation
**REFL & INV**

Input Reflection File Name: 20BHSD.ref.orig  ◆ New to SnB  ◇ Old SnB File

Input Invariant File Name: ◆ New  ◇ Existing

Date Resolution: Minimum 999.0 Maximum 3.00

Number of Reflections to use: 200  Number of Triples to Use: 2000

**TRIALS & CYCLES**

Starting Phases from: ◆ Random Atoms  ◇ Random Phases

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

Number of Trials: 1000  Starting at Trial: 1  Random Seed (Prime): 11909

Input Phase File Name:

Input Atom File Name:

Number of Shake-and-Bake Cycles: 10

Terminate trials failing the R-Ratio test? ◇ Yes  ◆ No  R-Ratio Cutoff: 0.23

**PHASE REFINEMENT**

Method: ◆ Parameter Shift  ◇ Tangent Formula

Parameter Shift Options: Phase Shift: 90.0  Number of Shifts: 2

Number of passes through phase set: 3

Tangent Formula Options: Number of passes through phase set:
CONSTRAINTS

Number Of Peaks To Select: 4

Fourier Grid Size (map resolution): 1.00  Minimum Interpeak Distance: 3.00

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

Number of special position peaks to keep: 0

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: job1

Keep complete trace file?  Yes  No

Number of SnB jobs to submit: 1 (processors available)

DISPLAY

Result files prefix: job1

Number of peaks to use: 0  Maximum bond distance (Å):

Number of large peaks to be distinguished: 0  Maximum bond distance (Å) for large peaks:

DISPLAY HISTOGRAM  VISUALIZE STRUCTURE

VIEW GEOMETRY LISTING