Ab Initio Phasing by Dual-Space Direct Methods

Charles M. Weeks[1], George M. Sheldrick[2], Russ Miller[1,3],
Isabel Usón[2] and Herbert A. Hauptman[1]

[1]Hauptman-Woodward Medical Research Institute, Inc.
73 High Street, Buffalo, NY 14203 USA
[2]Institut für Anorganische Chemie der Universität Göttingen, Tammannstrasse 4,
D-37077 Göttingen, Germany
[3]State University of New York at Buffalo
Dept. of Comp. Sci. & Eng., Buffalo, NY 14260 USA

Abstract

Shake-and-Bake is an *ab initio* direct method for solving the crystallographic phase problem. Its most distinctive feature is the repeated alternation of reciprocal-space phase refinement with a complementary real-space process that seeks to improve phases by applying constraints. The *Shake-and-Bake* philosophy has been implemented in two independent computer programs, *SnB* and SHELXD. These programs have proven capable of solving complete structures containing as many as 2000 independent non-H atoms provided that accurate diffraction data have been measured to a resolution of 1.2Å or better. By using anomalous difference data, solutions have also been obtained for substructures containing as many as 70 selenium atoms. Substructure data sets having a maximum resolution in the 2.25-5.0Å range have been used successfully.

1. Introduction

Ab initio methods for solving the crystallographic phase problem rely on diffraction amplitudes alone and do not require prior knowledge of any atomic positions. General features that are not specific to the structure in question (*e.g.*, the presence of disulfide bridges or solvent regions) can, however, be utilized. For the last three decades, most small-molecule structures have been routinely solved by *direct methods*, a class of *ab initio* methods in which probabilistic phase relations are used to derive reflection phases from the measured amplitudes. Direct methods, implemented in widely-used, highly-automated, computer programs such as MULTAN (Main *et al.*, 1980), SHELXS (Sheldrick, 1990), SAYTAN (Debaerdemaeker *et al.*, 1985) and SIR (Burla *et al.*, 1989), provide computationally efficient solutions for structures containing less than approximately 100 unique non-H atoms. However, larger structures are not consistently amenable to these programs. In fact, few unknown structures with more than 200 unique equal atoms have ever been solved using these programs. The *Shake-and-Bake* approach differs from conventional direct methods by repetitively and unconditionally *alternating* reciprocal-space phase refinement (*shaking*) with density modification (*baking*) to impose the phase constraints implicit in real space (Weeks *et al.*, 1993; Miller *et al.*, 1993). Consequently, it yields a computer-intensive algorithm, requiring two Fourier transformations during each cycle, which has been made feasible in recent years due to the tremendous increases in computer speed. The *Shake-and-Bake* philosophy provided the breakthrough needed to achieve automated direct-methods solutions for much larger structures than had been possible with conventional direct-methods programs.

1.1. Reciprocal-Space Phase Refinement or Expansion (*Shaking*)

Direct methods are based on the fact that there exist linear combinations of phases, called structure invariants, the values of which, in principle, depend only on the magnitudes of the *normalized structure factors*,

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

where $\mathbf{r}_j$ is the position vector of one of the N atoms, assumed identical, in the primitive unit cell. The most useful phase relationships are the three-phase or *triplet invariants*,

$$T_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \tag{2}$$

the most probable values of which are given by the conditional probability distribution

$$P(\Phi_{\mathbf{HK}}) = [2\pi I_0(A_{\mathbf{HK}})]^{-1}\ \exp(A_{\mathbf{HK}}\cos\Phi_{\mathbf{HK}}) \tag{3}$$

where

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

as illustrated in Figure 1 (Cochran, 1955). *Ab initio* phase determination by direct methods requires that individual phase values be derived from a set of triplets (*i.e.*, triplet invariants). In theory, any of a variety of optimization methods could be used to extract phase information. However, so far only two (tangent refinement and parameter-shift optimization of the minimal function) have been shown to be of practical value.

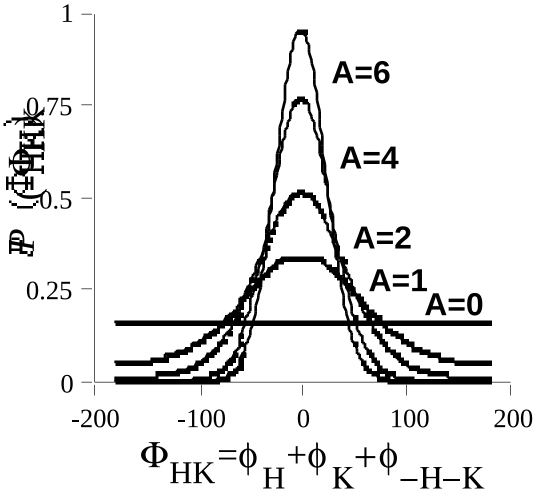


Figure 1. The conditional probability distribution of the three-phase structure invariants. Estimates of the invariant values are most reliable when the normalized structure-factor magnitudes ($|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$, and $|E_{\mathbf{-H-K}}|$) are large and the number of atoms in the unit cell, N, is small.

The Tangent Formula. If $\phi_{\mathbf{H}}$ is a new phase to be assigned, the *tangent formula*,

$$\tan\left(\phi_{\mathbf{H}}\right) = \frac{\sum_{\mathbf{K}} \left|E_{\mathbf{K}} E_{\mathbf{H-K}}\right| \sin\left(\phi_{\mathbf{K}} + \phi_{\mathbf{H-K}}\right)}{\sum_{\mathbf{K}} \left|E_{\mathbf{K}} E_{\mathbf{H-K}}\right| \cos\left(\phi_{\mathbf{K}} + \phi_{\mathbf{H-K}}\right)}, \qquad (5)$$

(Karle & Hauptman, 1956), provides the means used in conventional direct methods to compute $\phi_{\mathbf{H}}$. Furthermore, the tangent formula can also be used within the phase-refinement portion of the *Shake-and-Bake* procedure (Weeks *et al.*, 1994b; Sheldrick & Gould, 1995). The variance associated with $\phi_{\mathbf{H}}$ depends on $\Sigma_{\mathbf{K}} E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{-H-K}}/N^{1/2}$ and, in practice, the estimate is only reliable for $|E_{\mathbf{H}}|$>>1 and for structures with a limited number of atoms (N). If enough pairs of phases, $\phi_{\mathbf{K}}$ and $\phi_{\mathbf{-H-K}}$, are known, the tangent formula can be used to generate further phases ($\phi_{\mathbf{H}}$) which, in turn, can be combined with the observed amplitudes and included in the summation for subsequent reflections. Repeated iterations will permit most reflections with large $|E_{\mathbf{H}}|$ to be phased. If previously known phases are redetermined in each iteration, this process is one of *tangent-formula refinement;* if only new phases are determined, the phasing process is referred to as *tangent expansion.* When no initial phases are known, a ‘multisolution’ (Germain & Woolfson, 1968) or multi-trial approach is taken in which (*i*) a random-number generator is used to assign initial phase values (Baggio *et al.*, 1978; Yao, 1981), (*ii*) multiple sets of such trial phases are generated, and (*iii*) probable solutions are identified by ranking these phase sets according to suitable figures of merit.

The tangent formula can be derived using the assumption of equal resolved atoms. Nevertheless, it suffers from the disadvantage that, in space groups without translational symmetry, it is perfectly fulfilled by a false solution with all phases equal to zero, thereby giving rise to the so-called ‘uranium-atom’ solution with one dominant peak in the corresponding Fourier synthesis. In conventional direct-methods programs, the tangent formula is modified in various ways to include (explicitly or implicitly) information from the negative four-phase or quartet invariants (Schenk, 1974; Hauptman, 1974; Giacovazzo, 1976) that are based on the smallest as well as the largest E-magnitudes. Such modified tangent formulas do indeed largely overcome the problem of pseudosymmetric solutions for small N, but because the quartet term probabilities depend on 1/N, they are little more effective than the normal tangent formula for large N.

The Minimal Function. Constrained minimization of an objective function like the *minimal function*,

$$R(\Phi) = \sum_{\mathbf{H,K}} A_{\mathbf{HK}} \left[\cos T_{\mathbf{HK}} - \frac{I_1\left(A_{\mathbf{HK}}\right)}{I_0\left(A_{\mathbf{HK}}\right)} \right]^2 \Big/ \sum_{\mathbf{H,K}} A_{\mathbf{HK}} \qquad (6)$$

(Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.*, 1994) provides an alternative approach to phase refinement or phase expansion. $R(\Phi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and the expected values of the same triplets as given by the ratio of modified Bessel functions. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph (the minimal principle). Experimentation has thus far confirmed that, when the minimal function is used actively in the phasing process and solutions are produced, the final trial structure corresponding to the smallest value of $R(\Phi)$ is a solution provided that $R(\Phi)$ is calculated directly from the atomic positions before the phase-refinement step (Weeks *et al.*, 1994a). Therefore, $R(\Phi)$ is also an extremely useful figure of merit. The minimal function can also be written to include contributions from higher-order (*e.g.*, quartet) invariants although their use is not as imperative as with the tangent formula because the minimal function does not have a minimum when all phases are zero. In practice, quartets are rarely used in the minimal function because they increase the CPU time while adding little useful information for large structures because of the quartet probability dependence on $1/N$. The cosine function in Eq. 6 can also be replaced by other functions of the phases giving rise to alternative minimal functions. In particular, an exponential expression has been found to give superior results for several P1 structures (Hauptman *et. al.*, 1999).

Parameter Shift. In principle any minimization technique could be used to minimize $R(\Phi)$ by varying the phases. So far, a seemingly simple algorithm, known as parameter shift (Bhuiya & Stanley, 1963), has proven to be quite powerful and efficient as an optimization method when used within the *Shake-and-Bake* context to reduce the value of the minimal function. For example, a typical phase-refinement stage consists of three iterations or scans through the reflection list, with each phase being shifted a maximum of two times by 90° in either the positive or negative direction during each iteration. The refined value for each phase is selected, in turn, through a process which involves evaluating the minimal function using the original phase and each of its shifted values (Weeks *et al*., 1994a). The phase value that results in the lowest minimal-function value is chosen at each step. Refined phases are used immediately in the subsequent refinement of other phases. It should be noted that the parameter-shift routine is similar to that used in ψ-map refinement (White & Woolfson, 1975) and *XMY* (Debaerdemaeker & Woolfson, 1989).

1.2. Real-Space Constraints (*Baking*)

Peak picking is a simple but powerful way of imposing an atomicity constraint. The potential for real-space phase improvement in the context of small-molecule direct methods was recognized by Jerome Karle (1968). He found that even a relatively small, chemically-sensible fragment extracted by manual interpretation of an electron-density map could be parlayed into a complete solution by transformation back to reciprocal space and then performing additional iterations of phase refinement with the tangent formula. Automatic, real-space, electron-density map interpretation in the *Shake-and-Bake* procedure consists of selecting an appropriate number of the largest peaks in each cycle to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance between atoms. If markedly unequal atoms are present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space in a subsequent structure-factor calculation. 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. In theory, under appropriate conditions it should be possible to substitute alternative density-modification procedures such as low-density elimination (Shiono & Woolfson, 1992; Refaat & Woolfson, 1993) or solvent flattening (Wang, 1985), but no practical applications of such procedures have yet been made. The imposition of physical constraints counteracts the tendency of phase refinement to propagate errors or produce overly consistent phase sets. Several variants of peak picking, which are discussed below, have been successfully employed within the framework of *Shake-and-Bake*.

Simple Peak Picking. In its simplest form, peak picking consists of simply selecting the top N_u E-map peaks where N_u is the number of unique non-H atoms in the asymmetric unit. This is adequate for true small-molecule structures. It has also been shown to work well for heavy-atom or anomalously scattering substructures where N_u is taken to be the number of expected substructure atoms (Smith *et al.*, 1998; Turner *et al.*, 1998). For larger structures (N_u>100), it is likely to be better to select about $0.8N_u$ peaks, thereby taking into account the probable presence of some atoms that, owing to high thermal motion or disorder, will not be visible during the early stages of a structure determination. Furthermore, a recent study (Miller & Weeks, 1998) has shown that structures in the 250-1000 atom range which contain a half dozen or more moderately heavy atoms (*i.e.*, S, Cℓ, Fe) are more easily solved if only $0.4N_u$ peaks are selected. The only chemical information used at this stage is a minimum inter-peak distance, generally taken to be 1.0Å.

Iterative Peaklist Optimization. An alternative approach is to select approximately N_u peaks as potential atoms and then eliminate some of them, one by one, while maximizing a suitable figure of merit such as

$$P = \sum_{\mathbf{H}} \left| E_c^2 \right| \left(\left| E_o^2 \right| - 1 \right). \qquad (7)$$

The top N_u peaks are used as potential atoms to compute E_c. The atom that leaves the highest value of P is then eliminated. Typically, this procedure, which has been termed *iterative peaklist optimization* (Sheldrick & Gould, 1995), is repeated until only $2N_u/3$ atoms remain. Usage of Eq. 7 may be regarded as a reciprocal-space method of maximizing the fit to the origin-removed sharpened Patterson function, and it is used for this purpose in molecular replacement (Beurskens, 1981). Subject to various approximations, maximum likelihood considerations also indicate that it is an appropriate function to maximize (Bricogne, 1998). Iterative peaklist optimization provides a higher percentage of solutions than simple peak picking, but it suffers from the disadvantage of requiring much more CPU time.

Random Omit Maps. A third peak-picking strategy also involves selecting approximately N_u of the top peaks and eliminating some but, in this case, the deleted peaks are chosen at random (Sheldrick, 2000). Typically, 1/3 of the potential atoms are removed, and the remaining atoms are used to compute E_c. By analogy to the common practice in macromolecular crystallography of omitting part of a structure from a Fourier calculation in hopes of finding an improved position for the deleted fragment, this version of peak picking is described as making a *random omit map*. This procedure is a little faster than simply picking N_u atoms because fewer atoms are used in the structure-factor calculation. More important is the fact that, like iterative peaklist optimization, it has the potential for being a more efficient search algorithm.

1.3. Fourier Refinement (Twice Baking)

E-map recycling, but without phase refinement (Sheldrick, 1982, 1990; Kinneging & de Graaff, 1984), was frequently used in conventional direct-method programs to improve the completeness of the solutions after phase refinement. It is important to apply Fourier refinement to *Shake-and-Bake* solutions also because such processing significantly increases the number of resolved atoms, thereby making the job of map interpretation much easier. Since phase refinement *via* either the tangent formula or the minimal function requires relatively accurate invariants that can only be generated using the larger *E* magnitudes, a limited number of reflections is phased during the actual dual-space cycles. Working with a limited amount of data has the added advantage that less CPU time is required. However, if the current trial structure is the 'best' so far based on a figure of merit (either the minimal function or a real-space criterion), then it makes sense to subject this structure to Fourier refinement using additional data, thereby reducing series-termination errors. The correlation coefficient

$$CC = \left[\sum wE_o{}^2E_c{}^2 \cdot \sum w - \sum wE_o{}^2 \cdot \sum wE_c{}^2 \right] \Big/ \qquad (8)$$
$$\left\{ \left[\sum wE_o{}^4 \cdot \sum w - \left(\sum wE_o{}^2 \right)^2 \right] \cdot \left[\sum wE_c{}^4 \cdot \sum w - \left(\sum wE_c{}^2 \right)^2 \right] \right\}^{1/2}$$

(Fujinaga & Read, 1987), where *w* is a weight (usually unity), has been found to be an especially effective figure of merit when used with all the data and is, therefore, suited for identifying the most promising trial structure at the end of Fourier refinement. Either simple peak picking iterative peaklist optimization can be employed during the Fourier refinement cycles in conjunction with weighted *E*-maps (Sim, 1959). The final model can be further improved by isotropic displacement parameter (B_{iso}) refinement for the individual atoms (Usón *et al*., 1998), followed by calculation of the Sim (1959) or sigma-A (Read, 1986) weighted map. This is particularly useful when the requirement of atomic resolution is barely fulfilled, and it makes it easier to interpret the resulting maps by classical macromolecular methods.

2. Computer Programs (*SnB* and SHELXD)

The *Shake-and-Bake* algorithm has been implemented independently in two computer programs. These are (*i*) *SnB* written in Buffalo at the Hauptman-Woodward Institute, principally by Charles Weeks and Russ Miller (Miller *et al*., 1994; Weeks & Miller, 1999a), and (*ii*) SHELXD (which is also known by the alias 'Halfbaked'), written in Göttingen by George Sheldrick (Sheldrick, 1997, 1998). SHELXD attempts to do more during the real-space (*baking*) stage than is available to the user with the current version of *SnB*. The most recent public release of *SnB* is available on the Web at http://www.hwi.buffalo.edu/SnB/ along with documentation, test data, and other pertinent information. SHELXD will be released when testing is complete; for details see the SHELX homepage at http://shelx.uni-ac.gwdg.de/SHELX/.

2.1. Flowchart and Program Comparison

A flowchart for the generic *Shake-and-Bake* algorithm, which provides the foundation for both programs, is presented in Figure 2. It contains two refinement loops embedded in the trial structure loop. The first of these loops (steps 5-9) is a dual-space phase-improvement loop entered by all trial structures, and the second (steps 11-14) is a real-space Fourier-refinement

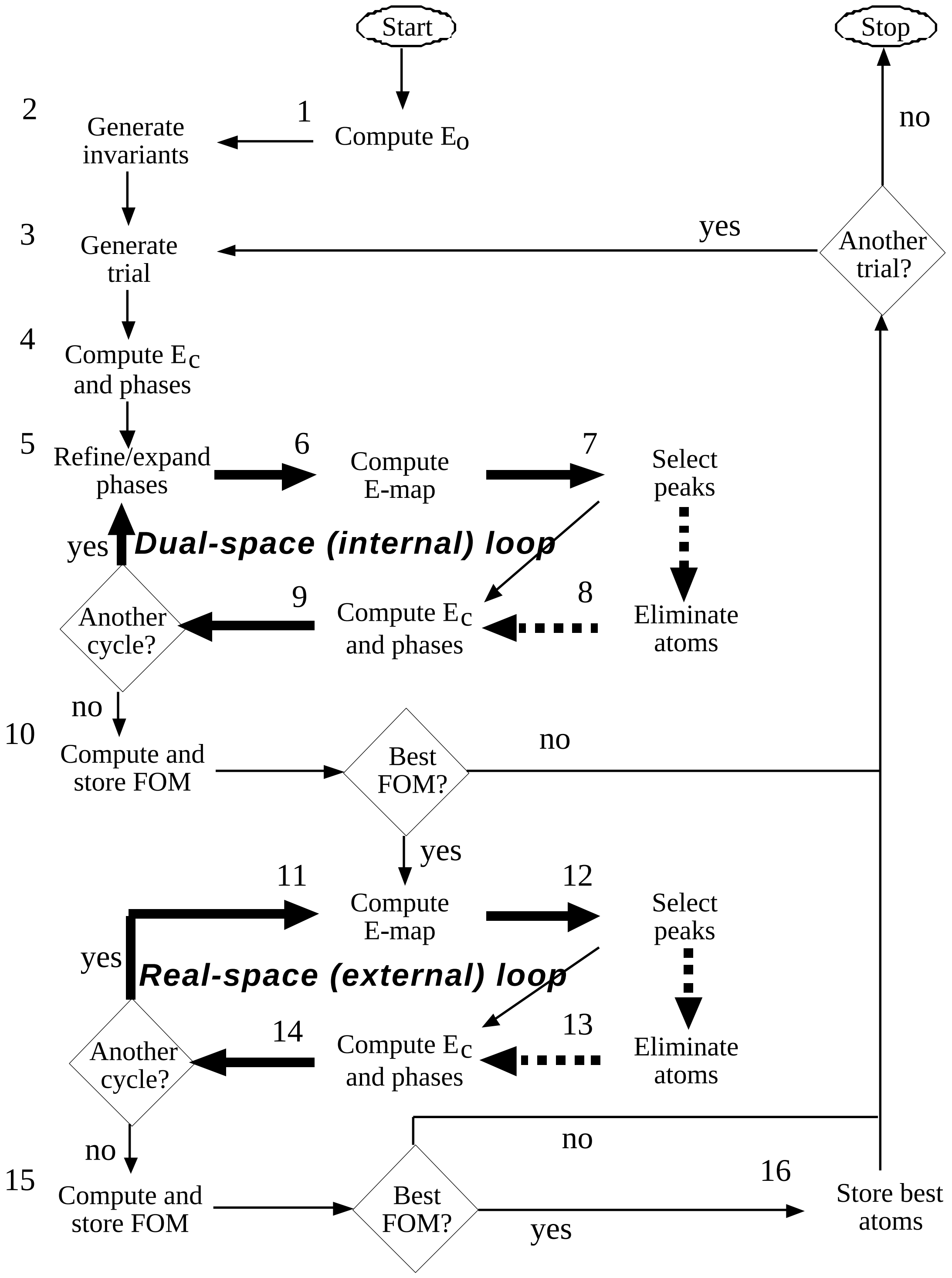


Figure 2. A flowchart for the *Shake-and-Bake* procedure, which is implemented in both *SnB* and SHELXD. The essence of the method is the dual-space approach of refining trial structures as they shuttle between real and reciprocal space. In the general case, steps 7 and 12 are any density-modification procedure, and steps 9 and 14 are inverse Fourier transforms rather than structure-factor calculations. The optional steps 8 and 13 take the form of *iterative peaklist optimization* or *random omit maps* in SHELXD. Any suitable starting model can be used in step 3, and SHELXD attempts to improve on random models (when possible) by utilizing Patterson-based information. Step 4 is bypassed if phase sets (random or otherwise) provide the starting point for the dual-space loop. SHELXD enters the real-space loop if the FOM (correlation coefficient) is within a specified threshold (1-5%) of the best value so far.

loop entered only by those trial structures that are currently judged to be the best on the basis of some figure of merit. These loops have been called the internal and external loops, respectively, in previous descriptions of the SHELXD program (*e.g.*, Sheldrick & Gould, 1995; Sheldrick, 1997, 1998). Currently, the major algorithmic differences between the programs are the following:

(a) During the reciprocal-space segment of the dual-space loop (Figure 2, step 5), *SnB* can perform tangent refinement or use parameter shift to reduce the minimal function (Eq. 6) or an exponential variant of the minimal function (Hauptman *et al*., 1999). SHELXD can perform either Karle-type tangent-expansion (Karle, 1968) or parameter-shift refinement based on either the minimal function or the tangent formula. During tangent or parameter-shift refinement, all phases computed in the preceding structure-factor calculation (step 4 or 9) are refined. During tangent expansion in SHELXD, the phases of (typically) the 40% highest calculated E-magnitudes are held fixed, and the phases of the remaining 60% are determined by using the tangent formula.

(b) In real space, *SnB* uses simple peak picking, varying the number of peaks selected on the basis of structure size and composition. SHELXD contains provisions for all the forms of peak picking described above.

(c) *SnB* relies primarily on the minimal function (Eq. 6) as a figure of merit whereas SHELXD uses the correlation coefficient (Eq. 8), calculated using all data, after the final dual-space (internal) cycle and in the real-space (external) loop.

2.2. Parameter Values

All of the major parameters of the *Shake-and-Bake* procedure (*i.e.*, the numbers of refinement cycles, phases, triplet invariant relationships, and peaks selected) are a function of structure size and can be expressed in terms of N_u, the number of unique non-H atoms in the asymmetric unit. These parameters have been fine-tuned in a series of tests using data for both small and large molecules (Weeks *et al.*, 1994a; Chang *et al.*, 1997; Miller & Weeks, 1998; Weeks & Miller, 1999b). Default parameter values used in the *SnB* program are summarized in Table 1. At resolutions in the 1.1-1.4Å range, recalcitrant data sets can sometimes be made to yield solutions if (*i*) the phase:invariant ratio is increased from 1:10 to values ranging between 1:20 and 1:50 or (*ii*) the number of dual-space refinement cycles is doubled or tripled. The presence of moderately heavy atoms (*e.g.*, S, Cℓ, Fe) greatly increases the probability of success at resolutions less than 1.2Å. Parameter recommendations for substructures are based on an analysis of the peak-wavelength anomalous difference data for S-adenosylhomocysteine (AdoHcy) hydrolase (Turner *et al.*, 1998). Parameter shift with a maximum of two 90° steps (indicated by the shorthand notation PS(90°,2)) is the default phase-refinement mode. However, some structures (especially large P1 structures) may respond better to a single larger shift (*e.g.*, PS(157.5°,1)) (Deacon *et al.*, 1998). This seems to reduce the frequency of false minima (see Section 4.2). In general, the parameter values used in SHELXD are similar to those used in *SnB*. However, the *combination* of random omit maps with tangent extension has been found to be the most effective strategy within the context of SHELXD. Consequently, it is used as the default.

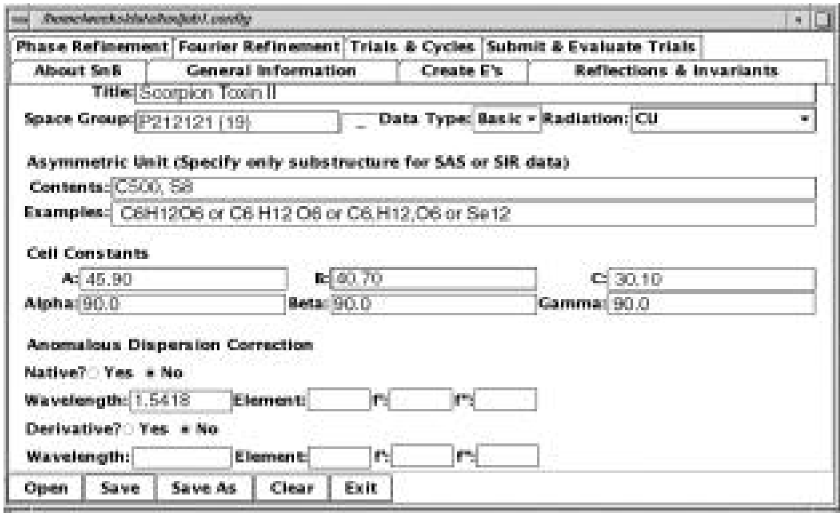
Table 1. Recommended parameter values for the *SnB* program expressed in terms of N_u, the number of unique non-H atoms (solvent atoms are typically ignored). Full structure recommendations are for data sets measured to 1.1Å resolution or better. Only heavy atoms or anomalous scatterers are counted for substructures.

Parameter	Full Structures	Substructures

Phases	$10N_u$	$30N_u$
Triplet Invariants	$100N_u$	$300N_u$
Peaks (with S, Cl)	$0.4N_u$	N_u
Peaks (no "heavy")	$0.8N_u$	
Cycles	$N_u/2$ if N_u<100 or if N_u<400 with S, $C\ell$, etc. N_u otherwise	$2N_u$

2.3. Using the Programs

On account of the intensive nature of the computations involved, *SnB* and SHELXD are designed to run unattended for long periods while also providing ways for the user to check the status of jobs in progress. The following brief description of *SnB* usage is provided as an example. The user interacts with the program via a graphical user interface (GUI). First, basic information about the unit cell and its contents along with the name of a reflection file containing either F or F^2 data are entered (Figure 3, top). Default values (which the user is free to change) are automatically supplied for most parameters following the guidelines presented in Table 1. *SnB* is linked to the DREAR package of data-processing routines (Blessing, Guo & Langs, 1996), which can then be used to generate normalized structure-factor magnitudes ($|E|$s) for traditional (full-structure) data sets as well as difference $|E|$s for SIR and SAS data sets. Currently, users wishing to base the phase determination on E_A must generate their own F_A values (Karle, 1980) and process them as non-difference data. After $|E|$s have been computed, the interface program can be used to submit the actual *Shake-and-Bake* job. The progress of on-going jobs can be followed by monitoring a figure-of-merit histogram for completed trial structures (Figure 3, right). A clear bimodal distribution of figure-of-merit values is a strong indication that a solution has, in fact, been found. However, not all solutions are so obvious, and it sometimes pays to inspect the best trial even when the histogram is unimodal. The course of a typical solution as a function of *Shake-and-Bake* cycle is contrasted to that of a nonsolution



Title: Scorpion Toxin II

Space Group: P212121 (19) Data Type: Basic Radiation: CU

Asymmetric Unit (Specify only substructure for SAS or SIR data)

Contents: C500, S8

Examples: C6H12O6 or C6 H12 O6 or C6,H12,O6 or Se12

Cell Constants

A: 45.90 B: 40.70 C: 30.10

Alpha: 90.0 Beta: 90.0 Gamma: 90.0

Anomalous Dispersion Correction

Native? ○ Yes ● No

Wavelength: 1.5418 Element: f': f":

Derivative? ○ Yes ● No

Wavelength: Element: f': f":

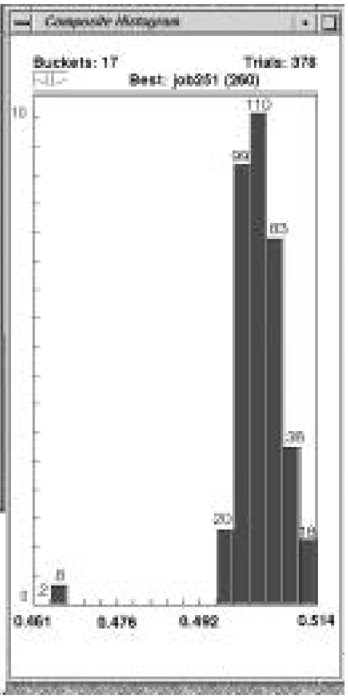


Figure 3 (above). A snapshot of the **General Information** page from the GUI interface for *SnB* v2.0. This screen contains all the necessary structure–specific information the user must supply. Selecting other pages allows all parameters to be examined and modified. The **Save** and **Open** functions allow the complete set of parameters to be stored in a "configuration" file and restored when needed. Figure 3 (right). A histogram of figure of merit values (minimal function) for 378 scorpion toxin II trials. This bimodal histogram suggests that 10 trials are solutions.

in Figure 4a. Minimal function values for a solution usually decrease abruptly over the course of just a few cycles, and a tool is provided within *SnB* that allows the user to visually inspect the trace of minimal function values for the best trial completed so far. Figure 4b shows that the abrupt decrease in minimal function values corresponds to a simultaneous abrupt increase in the number of peaks close to true atomic positions. In this example, a second abrupt increase in correct peaks occurs when Fourier refinement is started.

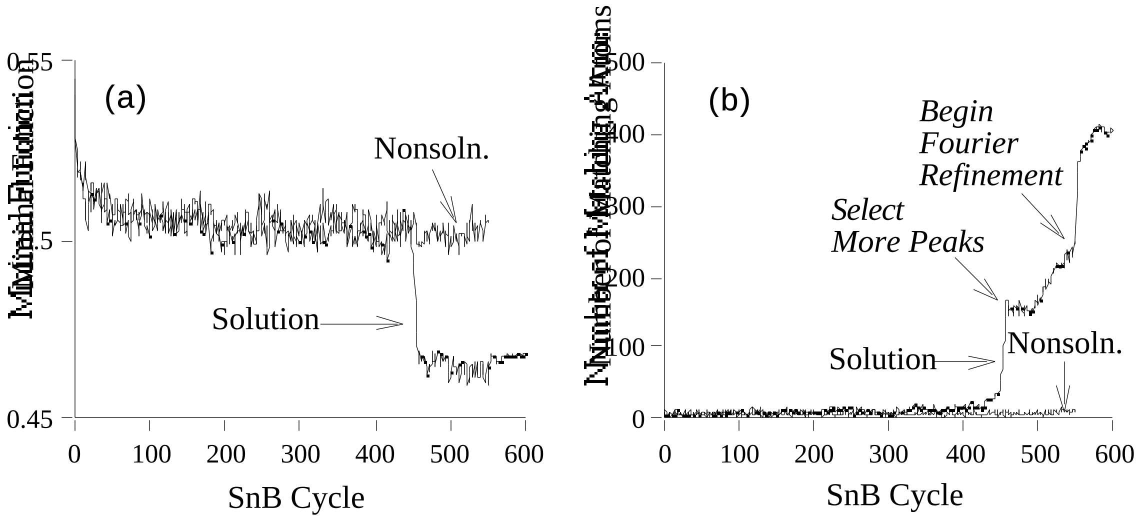


Figure 4. Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function of *Shake-and-Bake* cycle. (a) Minimal-function figure of merit, and (b) number of peaks closer than 0.5Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 500 (N_u) cycles, and 500 peaks (N_u) were then selected for an additional 50 ($0.1N_u$) dual-space cycles. The solution (which had the lowest minimal-function value) was then subjected to 50 cycles of Fourier refinement.

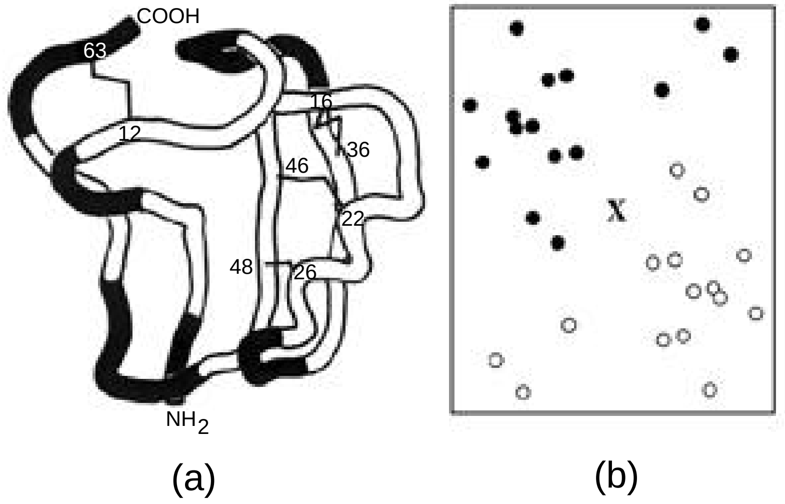


Figure 5. The high quality of *Shake-and-Bake* solutions is illustrated by (a) 47 of 64 residues traceable in the *SnB* solution of scorpion toxin II (Smith *et al.*, 1997) (Diagram courtesy of S. Ealick) and (b) the 30 selenium positions in the *SnB* solution of AdoHcy hydrolase (Turner *et al.*, 1998) viewed down the non-crystallographic two-fold axis (Diagram courtesy of P.L. Howell).

3. Applications

The solution of the (known) structure of triclinic lysozyme by SHELXD and shortly afterwards by *SnB* (Deacon *et al.*, 1998) finally broke the 1000-atom barrier for direct methods (there happen to be 1001 protein atoms in this structure!). Both programs have also solved a large number of previously unsolved structures that had defeated conventional direct methods; some examples are listed in Table 2. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier refinement stage. Two examples are shown in Figure 5.

4. **Discussion**

Most of the time, the *Shake-and-Bake* method works remarkably well, even for rather large structures. However, in problematic situations, the user needs to be aware of options that can increase the chance for success. The following discussion focuses on issues such as getting better initial trial structures, avoiding false minima, using *SnB* or SHELXD efficiently, and special considerations involved in the handling of substructures.

4.1. Getting a Better Start

When slightly heavier atoms such as sulfur are present, it is possible to start the *Shake-and-Bake* recycling procedure from a set of atomic positions that are consistent with the Patterson

Table 2. Some large structures solved by the *Shake-and-Bake* method. (a) Full structures (>300 atoms). (b) Se substructures (>25 Se) solved using peak-wavelength anomalous-difference data. Previously known test data sets are indicated by an asterisk (*). When two numbers are given in the resolution column, the second indicates the lowest resolution at which truncated data have yielded a solution. The program codes are *SnB* (S) and SHELXD (D).

	Space	N_u	N_u	N_u			
(a) Compound	Group	mol	+solv	heavy	d(Å)	Prog.	Ref.
Vancomycin	$P4_32_12$	202	258	8Cℓ	0.9-1.4	S	1
			312	6Cℓ	1.09	D	2
Actinomycin X2	P1	273	305	---	0.90	D	3
Actinomycin Z3	$P2_12_12_1$	186	307	2Cℓ	0.96	D	4
Actinomycin D	P1	270	314	---	0.94	D	4
Gramicidin A*	$P2_12_12_1$	272	317	---	0.86-1.1	S,D	5
DMSO d6 Peptide	P1	320	326	---	1.20	S	6
Er-1 Pheromone	C2	303	328	7S	1.00	S	7
Ristocetin A	$P2_1$	294	420	---	1.03	D	8
Crambin*	$P2_1$	327	423	6S	0.83-1.2	S,D	9,10
Hirustasin	$P4_32_12$	402	467	10S	1.2-1.55	D	11
Cyclodex. deriv.	$P2_1$	448	467	---	0.88	D	12
Alpha-1 Peptide	P1	408	471	Cℓ	0.92	S	13
Rubredoxin*	$P2_1$	395	497	Fe, 6S	1.0-1.1	S,D	14
Vancomycin	P1	404	547	12Cℓ	0.97	S	15
BPTI*	$P2_12_12_1$	453	561	7S	1.08	D	16
Cyclodex. deriv.	$P2_1$	504	562	28S	1.00	D	17
Balhimycin*	$P2_1$	408	598	8Cℓ	0.96	D	18
Mg-Complex*	$P\bar{1}$	576	608	8Mg	0.87	D	19
Scorpion Toxin II*	$P2_12_12_1$	508	624	8S	0.96-1.2	S	20
Amylose-CA26	P1	624	771	---	1.10	D	21
Mersacidin	$P3_2$	750	826	24S	1.04	D	22
Cv HiPIP H42Q*	$P2_12_12_1$	631	837	4Fe	0.93	D	23
HEW Lysozyme*	P1	1001	1295	10S	0.85	S,D	24,25
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20	D	23
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20	D	26

	Space	Mol. Wt.	Se	Se			
(b) Protein	Group	(kDa)	Located	Total	d(Å)	Prog.	Ref.
SAM Decarboxylase	$P2_1$	77	20	26	2.25	S	27
AIR Synthetase	$P2_12_12_1$	147	28	28	3.0	S	28
FTHFS	R32	200	28	28	2.5	D	29
AdoHcy Hydrolase	C222	95	30	30	2.8-5.0	S	30
Epimerase	$P2_1$	370	64	70	3.0	S	31

References: (1) Loll *et al.*, 1997; (2) Schäfer *et al.*, 1996; (3) Schäfer, 1998; (4) Schäfer *et al.*, 1998a; (5) Langs, 1988; (6) Drouin *et al*., 1998; (7) Anderson *et al*., 1996; (8) Schäfer & Prange, 1998; (9) Stec *et al.*, 1995; (10) Weeks *et al.*, 1995; (11) Usón *et al.*, 1999; (12) Aree *et al.*, 1999; (13) Prive *et al.*, 1999; (14) Dauter *et al.*, 1992; (15) Loll *et al.*, 1998; (16) Schneider, 1998; (17) Reibenspies, 1998; (18) Schäfer *et al.*, 1998b; (19) Teichert, 1998; (20) Smith *et al.*, 1997; (21) Gessler *et al.*, 1999; (22) Schneider *et al.*, 2000; (23) Parisini *et al.*, 1999; (24) Deacon *et al*., 1998; (25) Walsh *et al*., 1998; (26) Frazão *et al.*, 1999; (27) Ekstrom *et al.*, 1998; (28) Li *et al*., 1998; (29) Radfar *et al*., 2000; (30) Turner *et al*., 1998; (31) Deacon *et. al*., 2000.

function. For large structures, the vectors between such atoms will correspond to Patterson densities around or even below the noise level, so classical methods of locating the positions of these atoms unambiguously from the Patterson are unlikely to succeed. Nevertheless, the Patterson function can still be used to filter sets of starting atoms. This filter is currently implemented as follows in SHELXD. First, a sharpened Patterson function (Sheldrick *et al.*, 1993) is calculated, and the top (say) 200 non-Harker peaks further than a given minimum distance from the origin are selected, in turn, as two-atom translation search fragments, one such fragment being employed per solution attempt. For each of a large number of random translations, all unique Patterson vectors involving the two atoms and their symmetry equivalents are found and sorted in order of increasing Patterson density. The sum of the smallest 1/3 of these values is used as a figure of merit (PMF). Tests showed that, although the globally highest PMF for a given two-atom search fragment may not correspond to correct atomic positions, nevertheless by limiting the number of trials some correct solutions may still be found. After all the vectors have been used as search fragments (*e.g.,* after 200 attempts), the procedure is repeated starting again with the first vector. The two atoms may be used to generate further atoms using a full Patterson superposition minimum function or a weighted difference synthesis (in the current version of SHELXD, a combination of the two is used).

Table 3**.** Overall success rates for full structure solution for hirustasin using different two-atom search vectors chosen from the Patterson peak list.

Resolution (Å)	2-atom search fragments	Solutions per 1000 attempts
1.2	Top 100 general Patterson peaks	86
1.2	Top 300 general Patterson peaks	38
1.2	One vector, error = 0.08Å	14
1.2	One vector, error = 0.38Å	41
1.2	One vector, error = 0.40Å	219
1.2	One vector, error = 1.69Å	51
1.4	Top 100 general Patterson peaks	10
1.5	Top 100 general Patterson peaks	4
1.5	One vector, error = 0.29Å	61

In the case of the small protein BPTI (Schneider, 1998), 15300 attempts based on 100 different search vectors led to four final solutions with mean phase error less than 18˚ although none of the globally highest PMF values for any of the search vectors corresponded to correct solutions. Table 3 shows the effect of using different two-atom search fragments for hirustasin, a previously unsolved 55 amino-acid protein containing five disulfide bridges first solved using SHELXD (Usón et al., 1999). It is not clear why some search fragments perform so much better than others; surprisingly, one of the more effective search vectors deviates considerably (1.69Å) from the nearest true S-S vector.

4.2. Avoiding False Minima

The frequent imposition of real-space constraints appears to keep dual-space methods from producing most of the false minima that plague practitioners of conventional direct methods. Translated molecules have not been observed (so far), and traditionally problematical structures with polycyclic ring systems and long aliphatic chains are readily solved (McCourt *et al.*, 1996; McCourt *et al.*, 1997). False minima, of the type that occur primarily in space groups lacking translational symmetry and are characterized by a single large ‘uranium’ peak, do occur frequently in P1 and occasionally in other space groups. Triclinic hen egg-white lysozyme exhibits this phenomenon regardless of whether parameter-shift or tangent-formula phase refinement is employed (Deacon *et al.*, 1998). An example from another space group (C222) is provided by the Se substructure data for AdoHcy hydrolase (Turner *et al*., 1998). In this case, many trials converge to false minima if the feature in the *SnB* program that eliminates peaks at special positions is not utilized.

The problem with false minima is most serious if they have a ‘better’ value of the figure of merit being used for diagnostic purposes than do the true solutions. Fortunately, this is not the case with the uranium ‘solutions’, which can be distinguished on the basis of the minimal function (Eq. 6) or the correlation coefficient (Eq. 8). However, it would be inefficient to compute the latter in each dual-space cycle since it requires that essentially all reflections be used. To be an effective discriminator, the figure of merit must be computed using the phases calculated from the point-atom model, not from the phases directly after refinement. Phase refinement can and does produce sets of phases, such as the uranium phases, which do not correspond to physical reality. Hence, it should not be surprising that such phase sets might appear ‘better’ than the true phases and could lead to an erroneous choice for the best trial. Peak picking, followed by a structure-factor calculation in which the peaks are sensibly weighted, converts the phase set back to physically allowed values. If the value of the minimal function computed from the refined or *unconstrained* phases is denoted by R_{unc} and the value of the minimal function computed using the *constrained* phases resulting from the atomic model is denoted by R_{con}, then a function defined by

$$\text{R-Ratio} = (R_{\mathrm{con}} - R_{\mathrm{unc}}) / (R_{\mathrm{con}} + R_{\mathrm{unc}}) \qquad (9)$$

can be used to distinguish false minima from other nonsolutions as well as the true solutions. This distinction is illustrated for triclinic lysozyme in Figure 6a, and it is made possible by the fact that R_{unc} values are much smaller than normal for false minima. Once a trial falls into a false minimum, it never escapes. Therefore, the R-Ratio can be used, within *SnB*, as a criterion for early termination of unproductive trials. Based on data for several P1 structures, it appears that termination of trials with R-Ratio values exceeding 0.2 will eliminate most false minima without risking rejection of any potential solutions. In the case of triclinic lysozyme, false minima can be recognized, on the average, by cycle 25. Since the default recommendation would be for 1000 cycles, a substantial savings in CPU time is realized by using the R-Ratio early termination test. It should be noted that SHELXD optionally allows early termination of trials if the second peak is less than a specified fraction (*e.g.*, 40%) of the height of the first.

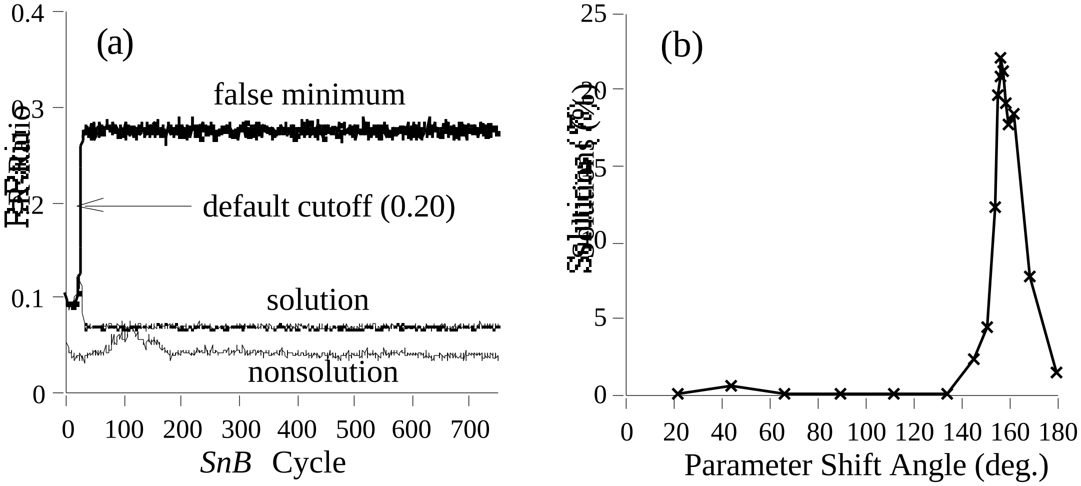


Figure 6. (a) R-Ratio values for triclinic lysozyme trials. (b) Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

Recognizing false minima is, of course, only part of the battle. It is also necessary to find a real solution, and essentially 100% of the triclinic lysozyme trials were found to be false minima when the standard parameter-shift conditions of two 90° shifts were used. In fact, significant numbers of solutions occur only when single-shift angles in the range 140-170° are used (Figure 6b), and there is a surprisingly high *success rate* (percentage of trial structures that go to solutions) over a narrow range of angles centered about 157.5°. It is also not surprising that there is a correlated decrease in the percentage of false minima in the range 140-150°. This suggests that a fruitful strategy for structures that exhibit a large percentage of false minima (*i.e.*, R-Ratio > 0.2), would be to run 100 or so trials at each of several shift angles in the range 90-180°, find the smallest angle which gives nearly zero false minima, and then use this angle as a single shift for many trials. This assumes, of course, that a solution is not already found while varying the shift angle. Balhimycin is an example of a large non-P1 structure that also requires a parameter shift of around 154° to obtain a solution using the minimal function.

4.3. Importance of Resolution and Complete Data

The importance of the presence of several atoms heavier than oxygen for increasing the chance of obtaining a solution by *SnB* at resolutions less than 1.2Å was noticed for truncated data from vancomycin and a 289-atom peptide structure crystallizing in space group I4 (Miller & Weeks, 1998). The results of SHELXD application to hirustasin (Usón *et al.*, 1998) are consistent with this. The 55 amino-acid protein hirustasin could be solved by SHELXD using either 1.2Å low-temperature data or 1.4Å room-temperature data; however, as shown in Figure 7a, the mean phase error (MPE) is significantly better for the 1.2Å data over the whole resolution range. The MPE is determined primarily by the data-to-parameter ratio, which is reflected in the smaller number of reliable triplet invariants at lower resolution. Although small-molecule interpretation based on peak positions worked well for the 1.2Å solution (overall MPE = 18°), standard protein chain tracing was required for the 1.4Å solution (overall MPE = 26°). As is clear from the corresponding electron-density map (Figure 7b), the *Shake-and-Bake* procedure produces easily interpreted protein density even when bonded atoms are barely resolved from each other. The hirustasin structure was also determined with SHELXD using 1.55Å truncated data, and this endeavor currently holds the record for the lowest-resolution successful application of *Shake-and-Bake* .

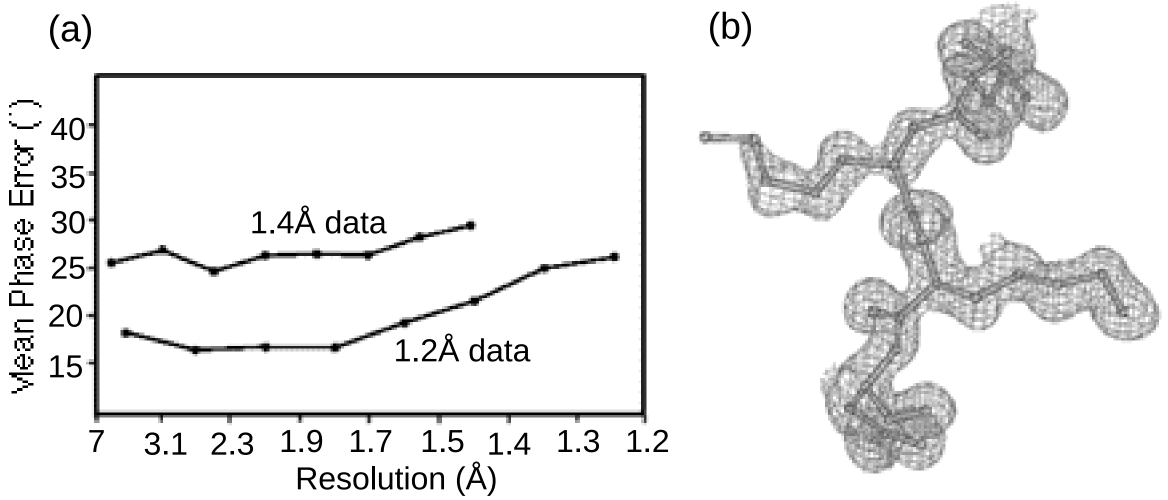


Figure 7. (a) Mean phase error as a function of resolution for the two independent *ab initio* SHELXD solutions of the previously unsolved protein hirustasin. Either the 1.2Å or the 1.4Å native data set led to solution of the structure. (b) Part of the hirustasin molecule from the 1.4Å room-temperature data after one round of B-value refinement with fixed coordinates.

The relative effects of accuracy, completeness, and resolution on *Shake-and-Bake* success rates using *SnB* for three large P1 structures were studied by computing error-free data using the known atomic coordinates. The results of these studies, presented in Table 4, show that experimental error contributed nothing of consequence to the low success rates for vancomycin and lysozyme. However, completing the vancomycin data up to the maximum measured resolution of 0.97Å resulted in a substantial increase in success rate which was further improved to an astounding success rate of 80% when the data were expanded to 0.85Å.

On account of overload problems, the experimental vancomycin data did not include any data at 10Å resolution or lower. A total of 4000 reflections were phased in the dual-space loop in the process of solving this structure with the experimental data. Some of these data were then replaced with the largest error-free magnitudes chosen from the missing reflections at several different resolution limits. The results in Table 5 show a ten-fold increase in success rate when only 200 of the largest missing magnitudes were supplied, and it made no difference whether these reflections had a maximum resolution of 2.8Å or were chosen from the whole 0.97Å sphere. The moral of this story is that, *when collecting data for Shake-and-Bake, it pays to take a second pass using a shorter exposure to fill-in the low-resolution data.*

Table 4. Success rates for three P1 structures illustrate the importance of using complete data to the highest possible resolution.

	Vancomycin	Alpha-1	Lysozyme
Atoms	547	471	~1200
Completeness	80.2%	85.6%	68.3%
Resolution	0.97Å	0.90Å	0.85Å
Parameter Shift	112.5°, 1	90°, 2	90°, 2
Success Rates			
Experimental	0.25%	14%	0%
Error-Free	0.2	19	0
Error-Free Complete	14	29	0.8
Error-Free Complete Extended to 0.85Å	80	42	---

Table 5. Improving success rates by ‘completing’ the vancomycin data.

Error-Free						
Refl. Added	0	100	200	200	400	800
		(3.5Å)	(2.8Å)	(random)	(1.3Å)	(1.1Å)
Success Rate	0.25%	0.3%	2.1%	2.4%	8.2%	11.1%

4.4. Random Omit Maps

Variations in the computational details of the dual-space loop can make major differences in the efficacy of *SnB* and SHELXD. The recent discovery of the power of random omit maps is a good illustration of this fact. Several strategies were combined in SHELXD and applied to a 148-atom P1 test structure (Karle *et al.,* 1989) with the results shown in Figure 8. The CPU time requirements of parameter-shift (**PS**) and tangent-formula expansion (**TE**) are similar, both being slower than no phase refinement (**NR**). In real space, the random-omit-map strategy (**RO**) was slightly faster than simple peak picking (**PP**) because fewer atoms were used in the structure-factor calculations. Both of these procedures were much faster than iterative peaklist optimization (**PO**). The original SHELXD algorithm (**TE+PO**) performs quite well in comparison with the *SnB* algorithm (**PS+PP**) in terms of the percentage of correct solutions, but less well when the efficiency is compared in terms of CPU time per solution. However the surprising result of these tests was that two curves involving random omit maps (**PS+RO** and **TE+RO**), which had been calculated as reference curves, are much more effective than the other algorithms, and even more so in terms of CPU efficiency. Indeed these two runs appear to approach a 100% success rate as the number of cycles becomes large!

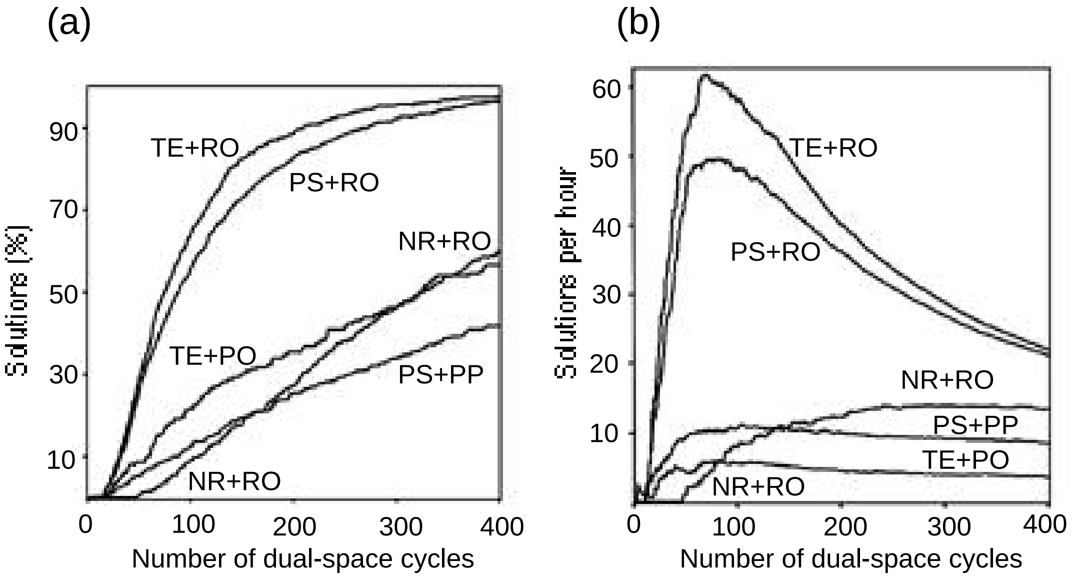
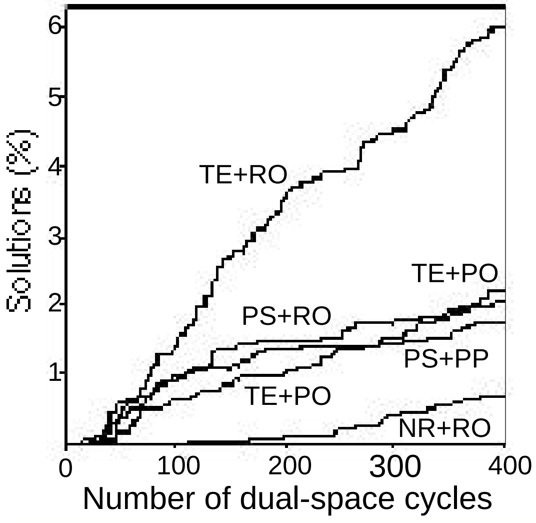


Figure 8. (a) Success rates and (b) cost effectiveness for several dual-space strategies as applied to a 148-atom P1 structure. Phase-refinement strategies: (**PS**) parameter-shift reduction of the minimal function value, (**TE**) Karle-type tangent expansion (holding the top 40% highest E_c fixed), and (**NR**) no phase refinement but Sim (1959) weights applied in the E-map (these depend on E_c and so cannot be employed after phase refinement). Real-space strategies: (**PP**) simple peak picking using $0.8N_u$ peaks, (**PO**) peaklist optimization (reducing N_u peaks to $2N_u/3$), and (**RO**) random omit maps (also reducing N_u peaks to $2N_u/3$). A total of about 10,000 trials of 400 internal loop cycles each were used to construct this diagram.

With hindsight, it is possible to understand why the random omit maps provide such an efficient *search algorithm.* In macromolecular structure refinement it is standard practice to omit parts of the model that do not fit well to the current electron density, to perform some refinement or simulated annealing (Hodel, Kim & Brünger, 1992) on the rest of the model to reduce memory effects, and then to calculate a new weighted electron-density map (omit map). If the original features reappear in the new density, they were probably correct; in other cases the omit map may enable a new and better interpretation. Thus, random omit maps should not lead to the loss of an essentially correct solution, but enable efficient searching in other cases.

Figure 9 illustrates the performance of the various strategies in the case of gramicidin A, a 317-atom structure that is arguably the most difficult structure to be solved by direct methods (Langs, 1988) prior to the introduction of the *Shake-and-Bake* philosophy. It should be noted that conventional direct methods incorporating the tangent formula tend to perform better in this space group ($P2_12_12_1$) than in P1, perhaps because there is less risk of a uranium-atom pseudosolution. Indeed, the combination of random omit maps and Karle-type tangent expansion in *Shake-and-Bake* is by far the most effective strategy for gramicidin A. In fact, tests using SHELXD on several structures have shown that the use of random omit maps is much more effective than picking the same final number of peaks from the top of the peak list. However, it should be stressed that it is the combination **TE+RO** that is particularly effective. A possible special case is when a very small number of atoms is sought (*e.g.*, Se atoms from MAD data). Preliminary tests indicate that peaklist optimization (**PO**) is competitive in such cases because the CPU time penalty associated with it is much smaller than it is when many atoms are involved. It is also interesting to note that the results presented in Figures 8 and 9 show that it is possible, albeit much less efficiently, to solve both structures using random omit maps without the use of any phase relationships based on probability theory (curve **NR+RO**)!

Figure 9. Percentage of correct solutions against cycle number for various strategies for the 317-atom structure, gramicidin A. A total of about 10,000 trials of 400 internal loop cycles each were used to construct this diagram.



4.5. Expand to P1?

The results shown in Table 5 and Figure 8 indicate that success rates in space group P1 can be anomalously high. This suggests that it might be advantageous to expand all structures to P1 and then to locate the symmetry elements afterwards. However, this is more computationally expensive than performing the whole procedure in the true space group, and in practice such a strategy is only competitive in low-symmetry space groups such as $P2_1$, C2 or P$\bar{1}$ (Chang, *et al*., 1997). Expansion to P1 also offers some opportunities for starting from 'slightly better than random' phases. One possibility, successfully demonstrated by Sheldrick & Gould (1995), is to use a rotation search for a small fragment (*e.g.*, a short piece of α-helix) to generate many sets of starting phases; after expansion to P1 the translational search usually required for molecular replacement is not needed. Various Patterson superposition minimum functions (Sheldrick & Gould, 1995; Pavelcik, 1994) can also provide an excellent start for phase determination for data expanded to P1. Drendel et al. (1995) were successful in solving small organic structures *ab initio* by a Fourier recycling method using data expanded to P1 without the use of probability theory.

4.6. Handling Substructures

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 isomorphous (Wilson, 1978) and anomalous (Mukherjee *et al*., 1989) difference structure factors. Experience has shown that successful substructure applications are highly dependent on the accuracy of the difference magnitudes. As the technology for producing selenomethionine-substituted proteins and collecting accurate multiple-wavelength (MAD) data (Hendrickson & Ogata, 1997; Smith, 1998) has improved, there has been increased interest in locating many selenium sites. For larger structures (say more than about 30 Se atoms), automated Patterson interpretation methods can be expected to run into difficulties since the number of unique peaks to be analyzed increases with the square of the number of atoms. Experimentally measured difference data is an approximation to the data for the hypothetical stubstructure, and it is reasonable to expect that conventional direct methods might run into difficulties sooner when applied to such data. Dual-space direct methods provide a more robust foundation for handling such data, which are often extremely noisy. Dual-space methods also have the added advantage that the expected number of Se atoms, N_u, which is usually known, can be exploited directly by picking the top N_u peaks. Successful applications require great care in data processing, especially if the F_A values resulting from a MAD experiment are to be used.

All successful applications of *SnB* to previously unknown SeMet data sets, as reported in Table 2, actually involved the use of peak-wavelength anomalous difference |*E*|s. The amount of data available for substructure 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. Guidelines for rejecting uncertain data have been suggested (Smith *et al.*, 1998), and it is essential that these guidelines 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.

On the other hand, it is also important that the phase:invariant ratio be maintained at 1:10 in order to ensure 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 triplets involving the largest reciprocal-lattice vectors may turn out to be too small. Consequently, a relatively small number of substructure phases (*e.g.*, $10N_u$) may not have a sufficient number (*i.e.*, $100N_u$) of invariants. Since the number of triplets increases rapidly with the number of reflections considered, the appropriate action in such cases is to increase the number of reflections as suggested in Table 1. This will typically produce the desired overdetermination.

It is rare for Se atoms to be closer to each other than 5Å, and the application of *SnB* to AdoHcy data (Turner *et al.*, 1998) truncated to 4Å and 5Å has been successful. Success rates were less for lower-resolution data, but the CPU time required per trial was also reduced, primarily because much smaller Fourier grids were necessary. Consequently, there was no net increase in the CPU time needed to find a solution.

A special version of SHELXD is being developed that makes extensive use of the Patterson function both in generating starting atoms and as a figure of merit. It has already successfully located the anomalous scatterers in a number of structures using MAD F_A data or simple anomalous differences. A recent example was the unexpected location of 17 anomalous scatterers (sulfur atoms and chloride ions) from the 1.5Å-wavelength anomalous differences of tetragonal HEW lysozyme (Dauter *et al.*, 1998).

5. Conclusions and Future Prospects

The *Shake-and-Bake* approach has increased, by an order of magnitude, the size of structures solvable by direct methods. Furthermore, a routine application of the *SnB* program to peak-wavelength anomalous difference data has revealed 64 of the 70 Se sites in a selenomethionine-substituted protein (Deacon *et al.*, 2000). Most importantly, there are clear indications that the method has not yet reached its limits.

The observations, reported above, of very high success rates for several sizeable P1 structures indicate that the full ramifications of *Shake-and-Bake* as an optimization method are not yet understood. The observation that a 317-atom structure could be solved without the use of any phase relationships based on probability theory (Figure 9, curve **NR+RO**) is also highly significant. There is no reason why a Fourier refinement technique, unlike probabilistic relations that become weaker as *N* increases, should not be applicable to very large structures. Larger structures require more computer time, but the amount of computer power available continues to increase. *SnB* and SHELXD are eminently parallelizable - they can simply be started with different random number seeds on all available CPU's! These programs can also be used in conjunction with software like CONDOR (Litzkow *et al.*, 1988) and GLOBUS (Foster & Kesselman, 1998) which increase the throughput of computer-intensive tasks by scavenging idle cycles on networks of PCs and workstations.

The requirement for data to very high resolution is, perhaps, more troublesome. One approach to extending these methods to lower resolution would be to replace the peak search by a search for small common fragments (*e.g.,* the five atoms of a peptide unit or an aromatic residue). This is also likely to be computationally intensive. It should also be possible to integrate the wARP procedure (Lamzin & Wilson, 1993; Perrakis *et al.*, 1997) into the real-space part of the *Shake-and-Bake* cycle. The Patterson function (Pavelcik, 1994; Sheldrick & Gould, 1995) and large Karle-Hauptman determinants (Vermin & de Graaff, 1978) might improve the success rate in borderline cases by providing better-than-random starting coordinates or phases.

Nevertheless, it is not necessarily true that peak picking is the primary limitation to lower-resolution applications. The lack of enough sufficiently accurate triplet invariant relationships may be a more fundamental problem. Simulation experiments have shown that, in theory, *SnB* can solve crambin even at 2.0Å if the invariants are accurate enough (Weeks, *et. al.*, 1998). Some of the underutilized formulas for invariant estimation that exist in the literature (*e.g.*, Hauptman, 1972) may be of some assistance in this regard. However, recent experimental work in the field of multiple-beam diffraction also provides grounds for hope. It has been shown that triplet invariants for a protein can be measured with a mean error of approximately 20° (Weckert *et al.*, 1993). In addition, direct methods strengthened by measured triplet invariants have been used to redetermine the structure of BPTI at resolutions as low as 2.0Å (Mathiesen & Mo, 1997, 1998). Currently, the one-at-a-time methods used to measure triplet phases seriously limit practical applications, but faster methods of data collection have been proposed (Shen, 1998). If the means can, in fact, be found for measuring significant numbers of triplet phases quickly and accurately, dual-space direct methods may become routinely applicable to much lower-resolution data than is currently possible.

6. Acknowledgements

The development, in Buffalo, of the *Shake-and-Bake* algorithm and the *SnB* program has been supported by grants GM-46733 from NIH and IRI-9412415 from NSF. The Buffalo authors would also like to thank the following individuals: (1) Chun-Shi Chang, Ashley Deacon, George DeTitta, Adam Fass, Steve Gallo, Hanif Khalak, Andrew Palumbo, Jan Pevzner, Thomas Tang, and Hongliang Xu who have aided the development of *SnB* and (2) Steve Ealick, P. Lynne Howell, Patrick Loll, Jennifer Martin, and Gil Prive who have generously supplied data sets. The development, in Göttingen, of SHELXD has been supported by HCM Institutional Grant ERB CHBG CT 940731 from the European Commission. The Göttingen authors wish to thank Thammarat Aree, Zbigniew Dauter, Judith Flippen-Anderson, Carlos Frazão, Jörg Kärcher, Katrin Gessler, Håkon Hope, Victor Lamzin, David Langs, Lukatz Lebioda, Paolo Lubini, Peer Mittl, Emilio Parisini, Erich Paulus, Ehmke Pohl, Thierry Prange, Joe Reibenspiess, Martina Schäfer, Thomas Schneider, Markus Teichert, László Vértesy, and Martin Walsh for discussions and/or generously providing data for structures referred to in this manuscript.

7. 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 raikovi*, *Acta Cryst*. **D52**, 469-480.

Aree, T., Usón, I., Schulz, B., Reck, G., Hoier, H., Sheldrick, G.M. & Saenger, W. (1999). Variation of a theme: crystal structure with four octakis (2,3,6-tri-O-methyl)-gamma-cyclodextrin molecules hydrated differently by a total of 19.3 water, *J. Am. Chem. Soc.* **121**, 3321-3327.

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 Cryst.* **A34**, 883-892.

Beurskens, P.T. (1981). A statistical interpretation of rotation and translation functions in reciprocal space, *Acta Cryst.* **A17**, 426-430.

Bhuiya, A.K. & Stanley, E. (1963). The refinement of atomic parameters by direct calculation of the minimum residual, *Acta Cryst.* **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 Cryst.* **D52**, 257-266.

Bricogne, G. (1998). Bayesian statistical viewpoint on structure determination: basic concepts and examples, *Meth. Enzymol.* **276**, 361-423.

Burla, M.C., Camalli, M., Cascarano, G., Giacovazzo, C., Polidori, G., Spagna, R. & Viterbo, D. (1989). SIR88 - a direct-methods program for the automatic solution of crystal structures, *J. Appl. Cryst.* **22**, 389-393.

Chang, C.-S., Weeks, C.M., Miller, R. & Hauptman, H.A. (1997). Incorporating tangent refinement in the *Shake-and-Bake* formalism, *Acta Cryst.*, **A53**, 436-444.

Cochran, W. (1955). Relations between the phases of structure factors, *Acta Cryst.*, **8**, 473-478.

Dauter, Z., Sieker, L.C. & Wilson, K.S. (1992). Refinement of rubredoxin from *Desulfovibrio vulgaris* at 1.0Å with and without restraints, *Acta Cryst.* **B48**, 42-59.

Dauter, Z., de la Fortelle, E. & Sheldrick, G.M. (1998). Phasing of lysozyme on anomalous sulfurs, Abstract 11.07.11, *Proc. Am. Cryst. Assoc. Meeting*, Arlington, VA.

Deacon, A.M., Weeks, C.M., Miller, R. & Ealick, S.E. (1998). The *Shake-and-Bake* structure determination of triclinic lysozyme, *Proc. Natl. Acad. Sci. USA*, **95**, 9284-9289.

Deacon, A.M., Ni, Y.S., Coleman Jr., W.G. & Ealick, S.E. (2000). The crystal structure of ADP-L-glycero-D-mannoheptose 6-epimerase: catalysis with a twist, *Structure* **8**, in press.

Debaerdemaeker, T. & Woolfson, M.M. (1983). On the application of phase relationships to complex structures XXII. Techniques for random phase refinement, *Acta Cryst.* **A39**, 193-196.

Debaerdemaeker, T., Tate, C. & Woolfson, M.M. (1985). On the application of phase relationships to complex structures. XXIV. The Sayre tangent formula, *Acta Cryst.* **A41**, 286-290.

Debaerdemaeker, T. & Woolfson, M.M. (1989). On the application of phase relationships to complex structures. XXVIII. *XMY* as a random approach to the phase problem, *Acta Cryst.* **A45**, 349-353.

DeTitta, G.T., Weeks, C.M., Thuman, P., Miller, R. & Hauptman, H.A. (1994). Structure solution by minimal function phase refinement and Fourier filtering: theoretical basis, *Acta Cryst*. **A50**, 203-210.

Drendel, W.B., Dave, R.D. & Jain, S. (1995). Forced coalescence phasing: a method for *ab initio* determination of crystallographic phases, *Proc. Natl. Acad. Sci. USA* **92**, 547-551.

Ekstrom, J.L., Stanley, B.A., Pegg, A.E. & Ealick, S.E. (1998). MAD structure of human S-adenosylmethionine decarboxylase, Abstract 11.02.05, *Proc. Am. Cryst. Assoc. Meeting*, Arlington, VA.

Foster, I. & Kesselman, C. (1998). The Globus project: a status report, *Proceedings of the Heterogeneous Computing Workshop*, IEEE Computer Society, J.K. Antonio (Ed.), pp. 4-18.

Frazão, C., Sieker, L., Sheldrick, G.M., Lamzin, V., LeGall, J. & Carrondo, M.A. (1999). *Ab initio* structure solution of a dimeric cytochrome c3 from *Desulfovibrio gigas* containing disulfide bridges, *J. Biol. Inorg. Chem.* **4**, 162-165.

Fujinaga, M. & Read, R.J. (1987). Experiences with a new translation-function program, *J. Appl. Cryst*. **20**, 517–521.

Germain, G. & Woolfson, M.M. (1968). On the application of phase relationships to complex structures, *Acta Cryst.* **B24**, 91-96.

Gessler, K., Usón, I., Takaha, T., Krauss, N., Smith, S.M., Okada, S., Sheldrick, G.M. & Saenger, W. (1999). V-Amylose at atomic resolution: x-ray structure of a cycloamylose with 26 glucoses, *Proc. Natl. Acad. Sci. USA* **96**, 4246-4251.

Giacovazzo, C. (1976). A probabilistic theory of the cosine invariant $\cos(\phi_{\mathbf{h}}+\phi_{\mathbf{k}}+\phi_{\mathbf{l}}-\phi_{\mathbf{-h-k-l}})$, *Acta Cryst,* **A32**, 91-99.

Hauptman, H. (1972). *Crystal Structure Determination: The Role of the Cosine Seminvariants*, New York: Plenum Press.

Hauptman, H.A. (1974). On the theory and estimation of the cosine invariants $\cos(\phi_{\mathbf{l}}+\phi_{\mathbf{m}}+\phi_{\mathbf{n}}+\phi_{\mathbf{p}})$, *Acta Cryst.* **A30**, 822-829.

Hauptman, H.A. (1991). A minimal principle in the phase problem, in *Crystallographic Computing 5: From Chemistry to Biology*, D. Moras, A.D. Podjarny & J.C. Thierry (Eds.), IUCr & Oxford University Press: Oxford, UK.

Hauptman, H.A., Xu, H., Weeks, C.M. & Miller, R. (1999). Exponential *Shake-and-Bake*: II. theoretical basis and applications, *Acta Cryst.* **A55**, 891-900.

Hendrickson, W.A. & Ogata, C.M. (1997). Phase determination from multiwavelength anomalous diffraction measurements, *Meth. Enzymol.* **276**, 494-523.

Hodel, A., Kim, S.-H. & Brünger, A.T. (1992). Model bias in macromolecular crystal structures, *Acta Cryst.* **A48**, 851-859.

Karle, I.L., Flippen-Anderson, J.L., Uma, K., Balaram, H. & Balaram, P. (1989). α-Helix and mixed 3_{10}/α-helix in cocrystallized conformers of Boc-Aib-Val-Aib-Aib-Val-Val-Val-Aib-Val-Aib-OMe, *Proc. Natl. Acad. Sci. USA* **86**, 765-769.

Karle, J. (1968). Partial structural information combined with the tangent formula for noncentrosymmetric crystals, *Acta Cryst.* **B24**, 182-186.

Karle, J. (1980). Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology, *Int. J. Quantum Chem. Quantum Biol. Symp.* **7**, 357-367.

Karle, J. & Hauptman, H. (1956). A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22, *Acta Cryst.* **9**, 635-651.

Kinneging, A.J. & de Graaff, R.A.G. (1984). On the automatic extension of incomplete models by iterative Fourier calculation, *J.Appl.Cryst.* **17**, 364-366.

Lamzin, V.S. & Wilson. K.S. (1993). Automatic refinement of protein models, *Acta Cryst.* **D49**, 129–147.

Langs, D.A. (1988). Three-dimensional structure at 0.86Å of the uncomplexed form of the transmembrane ion channel peptide gramicidin A, *Science* **241**, 188-191.

Li, C., Kappock, T.J., Stubbe, J. & Ealick, S.E. (1998). Personal communication.

Litzkow, M., Livny, M. & Mutka, M. (1988). Condor - a hunter of idle workstations, *Proc. 8th Intl. Conf. on Dist. Comp. Sys.*, 104-111.

Loll, P.J., Bevivino, A.E., Korty, B.D. & Axelsen, P.H. (1997). Simultaneous recognition of a carboxylate-containing ligand and an intramolecular surrogate ligand in the crystal structure of an asymmetric vancomycin dimer, *J. Am. Chem. Soc.* **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**, 293-298.

Main, P., Fiske, S.J., Hull, S.E., Lessinger, L., Germain, G., Declercq, J.P. & Woolfson, M.M. (1980). *MULTAN80: a system of computer programs for the automatic solution of crystal structures from x-ray diffraction data* (Universities of York and Louvain).

Mathiesen, R.H. and Mo, F. (1997). Application of known triplet phases in the crystallographic study of bovine pancreatic trypsin inhibitor. I: Studies at 1.55 and 1.75Å resolution, *Acta Cryst.* **D53**, 262-268.

Mathiesen, R.H. and Mo, F. (1998). Application of known triplet phases in the crystallographic study of bovine pancreatic trypsin inhibitor. II: Study at 2.0Å resolution, *Acta Cryst.* **D54**, 237-242.

McCourt, M.P., Li, N., Pangborn, W., Miller, R., Weeks, C.M. & Dorset, D.L. (1996). Crystallography of linear molecule binary solids. X-ray structure of a cholesteryl myristate/cholesteryl pentadecanoate solid solution, *J. Phys. Chem.* **100**, 9842-9847.

McCourt, M.P., Ashraf, K., Miller, R., Weeks, C.M., Li, N., Pangborn, W.A. & Dorset, D.L. (1997). X-ray crystal structures of cytotoxic oxidized cholesterols: 7-ketocholesterol and 25-hydroxycholesterol, *J. Lipid Research* **38**, 1014-1021.

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, J. Appl. Cryst.* **27**, 613–621.

Miller, R. & Weeks, C.M. (1998). *Shake-and-Bake*: applications and advances, *Direct Methods for Solving Macromolecular Structures*, S. Fortier (Ed.), Kluwer Academic Publishers: Dordrecht, pp. 389-400.

Mukherjee, A.K., Helliwell, J.R. & Main, P. (1989). The use of MULTAN to locate the positions of anomalous scatterers, *Acta Cryst.* **A45**, 715-718.

Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G.M. (1999). *Ab initio* solution and refinement of two high potential iron protein structures at atomic resolution, *Acta Cryst.* **D55**, 1773-1784.

Pavelcik, F. (1994). Patterson-oriented automatic structure determination. Deconvolution techniques in space group P1, *Acta Cryst.* **A50**, 467-474.

Perrakis, A., Sixma, T.A., Wilson, K.S. & Lamzin, V.S. (1997). wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models, *Acta Cryst.* **D53**, 448-455.

Prive, G., Anderson, D.H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). Packed protein bilayers in the 0.9_ resolution structure of a designed alpha helical bundle, *Protein Science* **8**, 1400-1409.

Radfar, R., Shin, R., Sheldrick, G.M., Minor, W., Lovell, C.R., Odom, J.D., Dunlap, R.B. & Lebioda, L. (2000). The crystal structure of N10-formyltetrahydrofolate synthetase from *Moorella thermoacetica*, *Biochemistry* **39**, in press.

Read, R.J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors, *Acta Cryst.* **A42**, 140–149.

Refaat, L.S. & Woolfson, M.M. (1993). Direct-space methods in phase extension and phase determination. II. Developments of low-density elimination, *Acta Cryst.* **D49**, 367-371.

Reibenspies, J. (1998). Personal communication.

Schäfer, M. (1998). Personal communication.

Schäfer, M. & Prange, T. (1998). Personal communication.

Schäfer, M., Schneider, T.R. & Sheldrick, G.M. (1996). Crystal structure of vancomycin, *Structure* **4**, 1509–1515.

Schäfer M., Sheldrick, G.M., Bahner, I. & Lackner, H. (1998a). Crystal structures of actinomycin D and Z3, *Angewandte Chemie*, **37**, 2381-2384.

Schäfer, M., Sheldrick, G.M., Schneider, T.R. & Vértesy, L. (1998b). Structure of balhimycin and its complex with solvent molecules, *Acta Cryst.* **D54**, 175-183.

Schenk, H. (1974). On the use of negative quartets, *Acta Cryst.* **A30**, 477-481.

Schneider, T.R. (1998). Personal communication.

Schneider, T.R., Kärcher, J., Pohl, E., Lubini, P., Vértesy, L. & Sheldrick, G.M. (2000). *Ab initio* crystal structure determination of the lantibiotic mersacidin, *Acta Cryst.* **D56**, in press.

Sheldrick, G.M. (1982). Crystallographic algorithms for mini- and maxi-computers, in *Crystallographic Computing,* D. Sayre (Ed.), Clarendon Press, Oxford, pp. 506–514.

Sheldrick, G.M. (1990). Phase annealing in SHELX-90: direct methods for larger structures, *Acta Cryst.* **A46**, 467–473.

Sheldrick, G.M. (1997). Direct methods based on real / reciprocal space iteration, in *Recent Advances in Phasing: Proceedings of the CCP4 Study Weekend* (DL-CONF-97-001), pp. 147-158, K.S. Wilson, G. Davies, A.S. Ashton, and S. Bailey (Eds.). CCLRC, Daresbury Laboratory, Warrington, UK.

Sheldrick, G.M. (1998). SHELX: Applications to macromolecules, *Direct Methods for Solving Macromolecular Structures*, S. Fortier (Ed.), Kluwer Academic Publishers: Dordrecht, pp. 401-411.

Sheldrick, G.M. (2000). In preparation.

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. & Sieker, L.C. (1993). The application of direct methods and Patterson interpretation to high-resolution native protein data, *Acta Cryst.* **D49**, 18–23.

Sheldrick, G.M. & Gould, R.O. (1995). Structure solution by iterative peaklist optimization and tangent expansion in space group P1, *Acta Cryst.* **B51**, 423–431.

Shen, Q. (1998). Solving the phase problem using reference-beam X-ray diffraction, *Physical Rev. Letters* **80**, 3268-3271.

Shiono, M. and Woolfson, M.M. (1992). Direct-space methods in phase extension and phase determination. I. Low-density elimination, *Acta Cryst.* **A48**, 451-456.

Sim, G.A. (1959). The distribution of phase angles for structures containing heavy atoms, *Acta Cryst.* **12**, 813-815.

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 determination and refinement of a scorpion protein toxin, *Acta Cryst.* **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 Cryst.* **D54**, 799-804 .

Smith, J.L. (1998). Multiwavelength anomalous diffraction in macromolecular crystallography, in *Direct Methods for Solving Macromolecular Structures*, S. Fortier (Ed.), Kluwer Academic Publishers: Dordrecht, pp. 211-225.

Stec, B., Zhou, R. & Teeter, M.M. (1995). Full-matrix refinement of the protein crambin at 0.83A and 130K, *Acta Cryst.* **D51**, 663-681.

Teichert, M. (1998). Personal communication.

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**, 369-375.

Usón, I., Sheldrick, G.M., de la Fortelle, E., Bricogne, G., di Marco, S., Priestle, J.P., Grütter, M.G. & Mittl, P.R.E. (1999). The 1.2Å crystal structure of hirustasin reveals the intrinsic flexibility of a family of highly disulphide bridged inhibitors, *Structure* **7**, 55-63.

Vermin, W.J. & de Graaff, R.A.G. (1978). The use of Karle-Hauptman determinants in small-structure determinations, *Acta Cryst.* **A34**, 892-894.

Walsh, M.A., Schneider, T.R., Sieker, L.C., Dauter, Z., Lamzin, V. & Wilson, K.S. (1998). Refinement of triclinic hen egg-white lysozyme at atomic resolution, *Acta Cryst.* **D54**, 522-546.

Wang, B.-C. (1985). Solvent flattening, *Meth. Enzymol.* **115**, 90-112.

Weckert, E., Schwegle, W. & Hümmer, K. (1993). Direct phasing of macromolecular structures by three-beam diffraction, *Proc. R. Soc. Lond. A*, **442**, 33-46.

Weeks, C.M., DeTitta, G.T., Miller, R. & Hauptman, H.A. (1993). Applications of the minimal principle to peptide structures, *Acta Cryst.* **D49**, 179-181.

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 Cryst.* **A50**, 210-220.

Weeks, C.M., Hauptman, H.A., Chang, C.-S. & Miller, R. (1994b). Structure determination by *Shake-and-Bake* with tangent refinement, *ACA Transactions Symposium* **30**, 153-161.

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 Cryst.* **D51**, 33-38.

Weeks, C.M., Miller R., and Hauptman, H.A. (1998), Extending the resolving power of *Shake-and-Bake*. In *Direct Methods for Solving Macromolecular Structures,* S. Fortier (Ed.), Kluwer Academic Publishers: Dordrecht, pp 463-468.

Weeks, C.M. & Miller, R. (1999a). The design and implementation of *SnB* v2.0, *J. Appl. Cryst.* **32**, 120-124.

Weeks, C. M. & Miller, R. (1999b). Optimizing *Shake-and-Bake* for proteins, *Acta Cryst.* **D55**, 492-500.

White, P.S. & Woolfson, M.M. (1975). The application of phase relationships to complex structures. VII. Magic integers, *Acta Cryst.* **A31**, 53-56.

Wilson, K.S. (1978). The application of MULTAN to the analysis of isomorphous derivatives in protein crystallography, *Acta Cryst.* **B34**, 1599-1608.

Yao, J.-X. (1981). On the application of phase relationships to complex structures. XVIII. RANTAN - random MULTAN, *Acta Cryst.* **A37**, 642-664.