16. DIRECT METHODS

Table 16.1.8.1
Overall success rates for full structure solution for hirustasin using different two-atom search vectors chosen from the Patterson peak list

Resolution (Å)	Two-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 = 0.08 Å	14
1.2	One vector, error = 0.38 Å	41
1.2	One vector, error = 0.40 Å	219
1.2	One vector, error = 1.69 Å	51
1.4	Top 100 general Patterson peaks	10
1.5	Top 100 general Patterson peaks	4
1.5	One vector, error = 0.29 Å	61

may be used to generate further atoms using a full Patterson superposition minimum function or a weighted difference synthesis.

In the case of the small protein BPTI (Schneider, 1998), 15 300 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 16.1.8.1 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 (1.69 Å) from the nearest true S–S vector.

16.1.9. *Shake-and-Bake*: an analysis of a dual-space method in action

The *Shake-and-Bake* algorithm generated the *SnB* program written in Buffalo at the Hauptman–Woodward Institute, principally by Charles Weeks and Russ Miller (Miller *et al.*, 1994; Weeks & Miller, 1999*a*). *SHELXD* (Usón & Sheldrick, 1999; Schneider & Sheldrick, 2002) and later *HySS* (Grosse-Kunstleve & Adams, 2003) were both inspired by *SnB* and employ the *Shake-and-Bake* philosophy with various modifications, in particular involving the use of the Patterson function to obtain starting phases. It is instructive to see how such software works in detail.

#### 16.1.9.1. *Flowchart and program comparison*

A flowchart for the generic *Shake-and-Bake* algorithm, which provides the foundation for these programs, is presented in Fig. 16.1.9.1. 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 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 (Fig. 16.1.9.1, step 5), *SnB* can perform tangent refinement or use parameter shift to reduce the minimal function [equation (16.1.4.2)] or an exponential variant of the minimal function (Hauptman *et al.*, 1999). *SHELXD* performs Karle-

type tangent expansion (Karle, 1968). During tangent or parameter-shift refinement with *SnB*, all phases computed in the preceding structure-factor calculation (steps 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. If there is a tendency for *SHELXD* to produce uranium-atom solutions, more phases should be held fixed in the tangent phase expansion.

(*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 [equation (16.1.4.2)] as a figure of merit whereas *SHELXD* uses the correlation coefficient [equation (16.1.6.1)], calculated using all data, after the final dual-space (internal) cycle and in the real-space (external) loop. In addition, *SHELXD* calculates a further correlation coefficient, CC_{weak}, calculated in the same
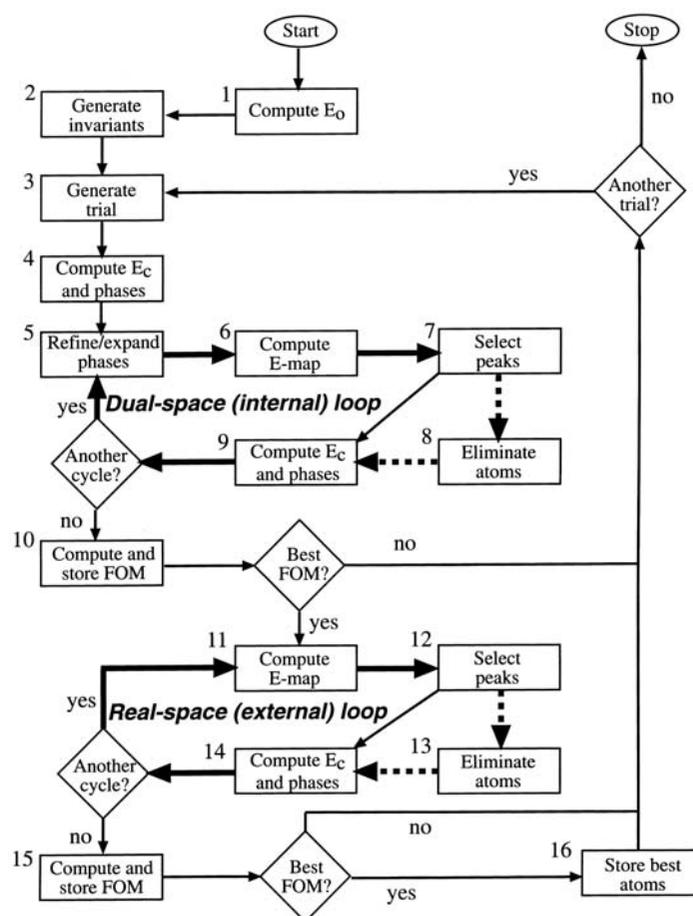
Figure 16.1.9.1
A flowchart for the *Shake-and-Bake* procedure, which is implemented in both *SnB* and *SHELXD*. The essence of the method is the dual-space approach of refining trial structures as they shuttle between real and reciprocal space. In the general case, steps 7 and 12 are any density-modification 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.

Table 16.1.9.1
Recommended parameter values for the *SnB* program

Values are 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 1.1 Å 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) Peaks (no 'heavy')	$0.4N_u$ $0.8N_u$	N_u
Cycles	$N_u/2$ if $N_u < 100$ or if $N_u < 400$ with S, Cl *etc.*; N_u otherwise	$2N_u$ (minimum 20)

way but using only the weak reflections, *i.e.* those not used directly for phasing.

16.1.9.2. Parameters and procedures

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, DeTitta *et al.*, 1994; Chang *et al.*, 1997; Weeks & Miller, 1999*b*). Default (recommended) parameter values used in the *SnB* program are summarized in Table 16.1.9.1. At resolutions in the 1.1–1.4 Å range, recalcitrant data sets can sometimes be made to yield solutions if (1) the phase:invariant ratio is increased from 1:10 to values ranging between 1:20 and 1:50 or (2) the number of dual-space refinement cycles is doubled or tripled. The presence of moderately heavy atoms (*e.g.* S, C, Fe) greatly increases the probability of success at resolutions less than 1.2 Å; in general, the higher the fraction of such atoms the more the resolution requirement can be relaxed, provided that these atoms have low B values. Thus, disulfide bridges are much more helpful than methionine sulfur atoms because they tend to have lower B values. 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 *P*1 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 16.1.10.1).

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* for *ab initio* solution of the full structure, and so is used as the default mode (see Section 16.1.10.2 for details). For substructure solution, especially for small substructures, it is normally faster to use starting atoms from Patterson seeding. Although both random omit and Patterson seeding can increase the success rate by an order of magnitude, combining both does not produce much further improvement. For very large substructures, and especially for very high symmetry space groups where the Patterson analysis is more time consuming, the random-omit procedure can be the more effective of the two. The largest substructure solved

by *SHELXD* is probably PDB code 2pnk (to be published), solved by Qingping Xu of the Joint Center for Structural Genomics (JCSG), with 197 correct and no incorrect Se sites out of 205 (the other eight were disordered). About 1.6 million trials were needed (using multiple CPUs) to obtain one correct solution when Patterson seeding was employed, but with the random-omit method many good solutions were obtained. This example also illustrates the point that it is important not to give up too soon!

The substructure solution program *HySS* in the *PHENIX* system is more-or-less a clone of *SHELXD*. For further details see Section 16.1.12.5.

16.1.9.3. Recognizing solutions

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 progress of current *SnB* jobs can be followed by monitoring a figure-of-merit histogram for the trial structures that have been processed (Fig. 16.1.9.2). 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 *SnB* cycle is contrasted with that of a nonsolution in Fig. 16.1.9.3. 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. Fig. 16.1.9.3 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
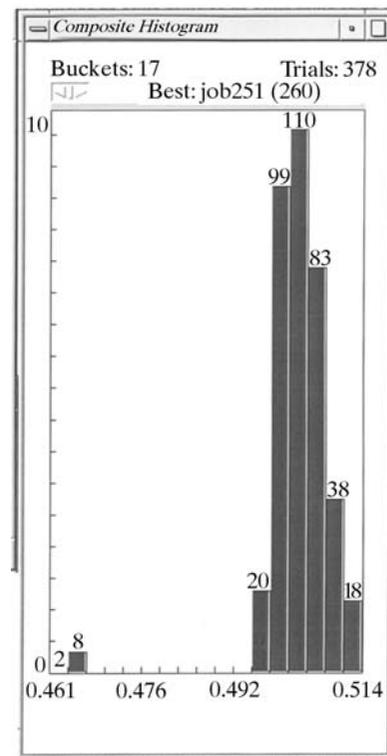
Figure 16.1.9.2
A histogram of figure-of-merit values (minimal function) for 378 scorpion toxin II trials. This bimodal histogram suggests that ten trials are solutions.
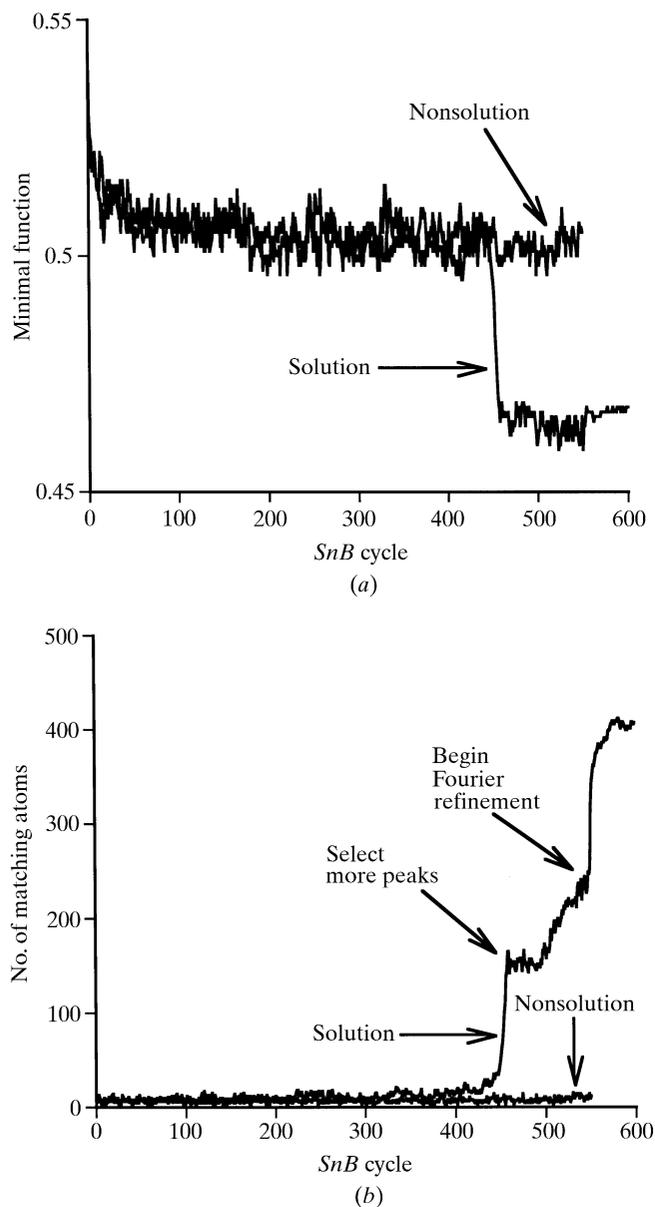
Figure 16.1.9.3
Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function *of Shake-and-Bake* cycle. (*a*) Minimal-function figure of merit, and (*b*) number of peaks closer than 0.5 Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 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.

this example, a second abrupt increase in correct peaks occurs when Fourier refinement is started.

Since the correlation coefficient is a relatively absolute figure of merit (given atomic resolution, values greater than 65% almost invariably correspond to correct solutions), it is usually clear when *SHELXD* has solved a structure, although when the data do not extend to atomic resolution the CC values are less informative, and for a substructure they depend strongly on the data quality.

16.1.10. Applying dual-space programs successfully

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 16.1.10.1. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier-refinement stage. 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 of success.

16.1.10.1. 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 problematic structures with polycyclic ring systems and long aliphatic chains are readily solved (McCourt *et al.*, 1996, 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 *P*1 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. An example from another space group (*C*222) 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 [equation (16.1.4.2)] or the correlation coefficient [equation (16.1.6.1)]. 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

$$R \text{ ratio} = (R_{con} - R_{unc})/(R_{con} + R_{unc}) \qquad (16.1.10.1)$$

can be used to distinguish false minima from other nonsolutions as well as the true solutions (Xu *et al.*, 2000). 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 *P*1 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 average, by cycle 25. Since the default recommendation would be for 1000 cycles, a substantial saving in CPU time is realized by using the *R* ratio early-termination test.

references