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.

16.1.7. Computer programs for dual-space phasing

The *Shake-and-Bake* algorithm has been implemented independently in two computer programs. These are (1) *SnB* written in Buffalo at the Hauptman–Woodward Institute, principally by Charles Weeks and Russ Miller (Miller *et al.*, 1994; Weeks & Miller, 1999*a*), and (2) *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 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/.

16.1.7.1. *Flowchart and program comparison*

A flowchart for the generic *Shake-and-Bake* algorithm, which provides the foundation for both programs, is presented in Fig. 16.1.7.1. 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. 16.1.7.1, 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* 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 [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.

16.1.7.2. *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

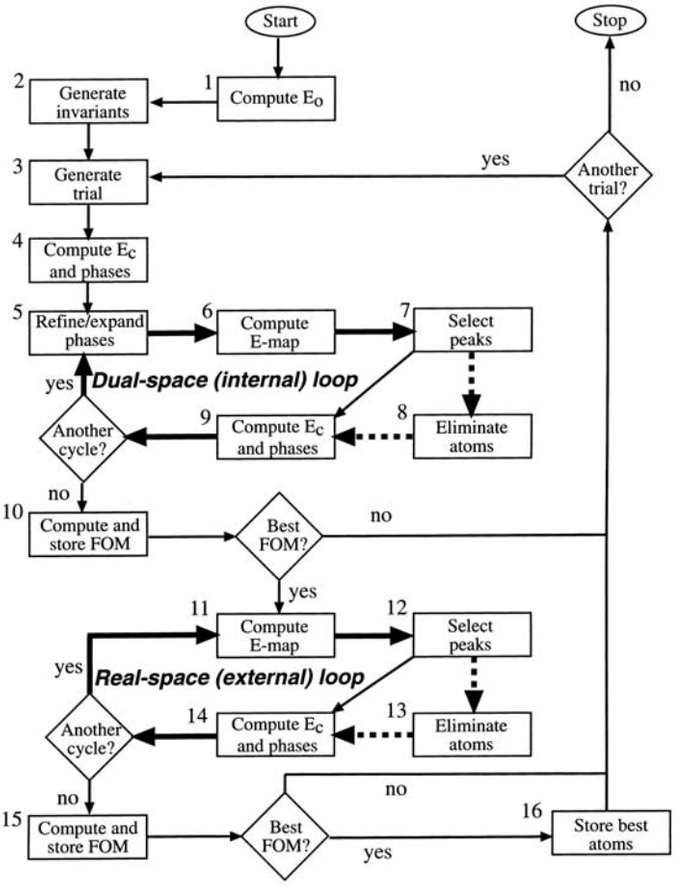

Fig. 16.1.7.1. A flowchart for the *Shake-and-Bake* procedure, which is implemented in both *SnB* and *SHELXD*. The essence of the method is the dual-space approach of refining trial structures as they shuttle between real and reciprocal space. In the general case, steps 7 and 12 are any density-modification 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.

are summarized in Table 16.1.7.1. 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 $P1$ structures) may respond better to a single larger shift [*e.g.* PS(157.5°, 1)]

Table 16.1.7.1. *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)

(Deacon *et al.*, 1998). This seems to reduce the frequency of false minima (see Section 16.1.8.2).

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 operational mode (see Section 16.1.8.4 for details).

16.1.7.3. *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. 16.1.7.2). 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. 16.1.7.3. 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. 16.1.7.3 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 this example, a second

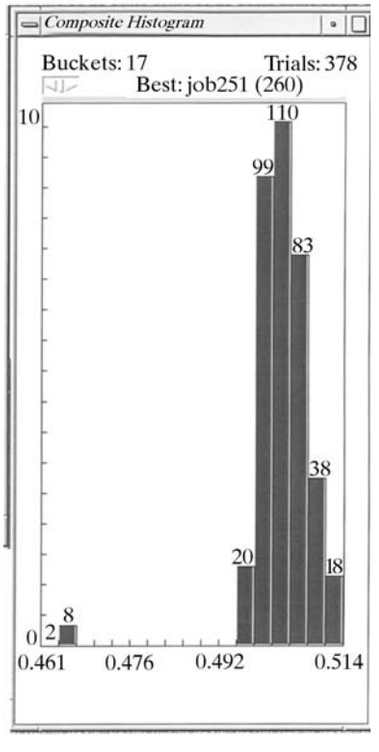

Fig. 16.1.7.2. A histogram of figure-of-merit values (minimal function) for 378 scorpion toxin II trials. This bimodal histogram suggests that ten trials are solutions.

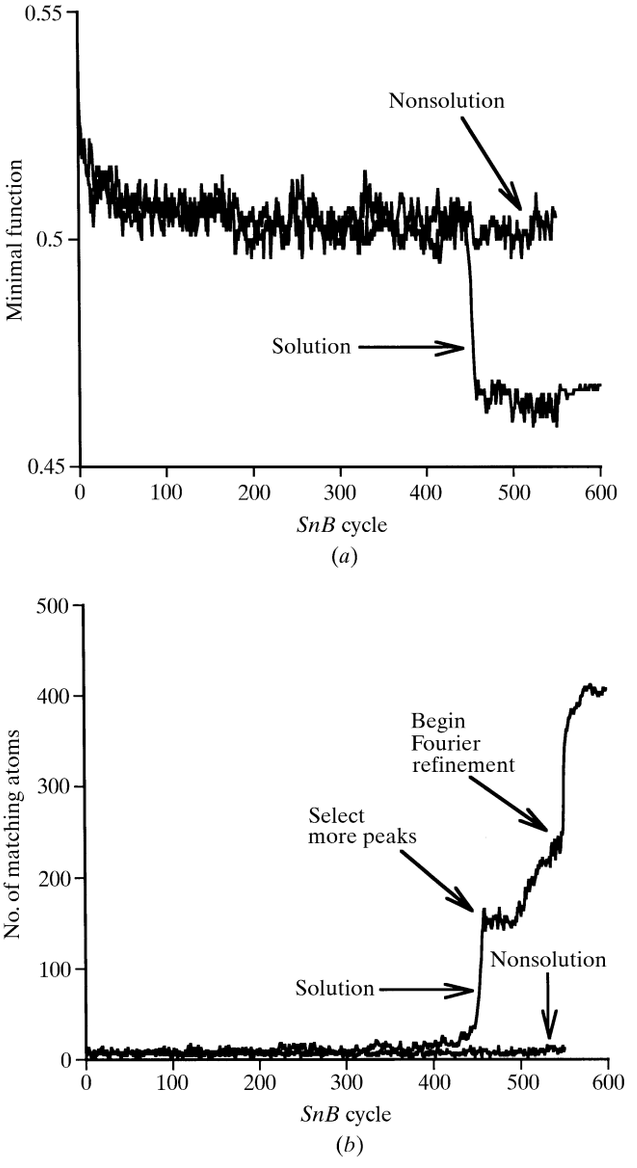

Fig. 16.1.7.3. Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function *of Shake-and-Bake* cycle. (*a*) Minimal-function figure of merit, and (*b*) number of peaks closer than 0.5 Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 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.

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. The current version of *SHELXD* includes an option for calculating it using the full data every 10 or 20 internal loop cycles, and jumping to the external loop if the value is high enough. Recalculating it every cycle would be computationally less efficient overall.

16.1.8. 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 16.1.8.1. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier-

Table 16.1.8.1. *Some large structures solved by the Shake-and-Bake method*

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).

(*a*) Full structures (>300 atoms).

Compound	Space group	N_u (molecule)	N_u + solvent	N_u (heavy)	Resolution (Å)	Program	Reference
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]
Cyclodextrin derivative	$P2_1$	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]
Cyclodextrin derivative	$P2_1$	504	562	28S	1.00	D	[17]
Balhimycin*	$P2_1$	408	598	8Cl	0.96	D	[18]
Mg-complex*	$P1$	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]

(*b*) Se substructures (> 25 Se) solved using peak-wavelength anomalous-difference data.

Protein	Space group	Molecular weight (kDa)	Se located	Se total	Resolution (Å)	Program	Reference
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, Sheldrick, Bahner & Lackner (1998); [5] Langs (1988); [6] Drouin (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] Reibenspiess (1998); [18] Schäfer, Sheldrick, Schneider & Vértesy (1998); [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.* (1999); [28] Li *et al.* (1999); [29] Radfar *et al.* (2000); [30] Turner *et al.* (1998); [31] Deacon & Ealick (1999).