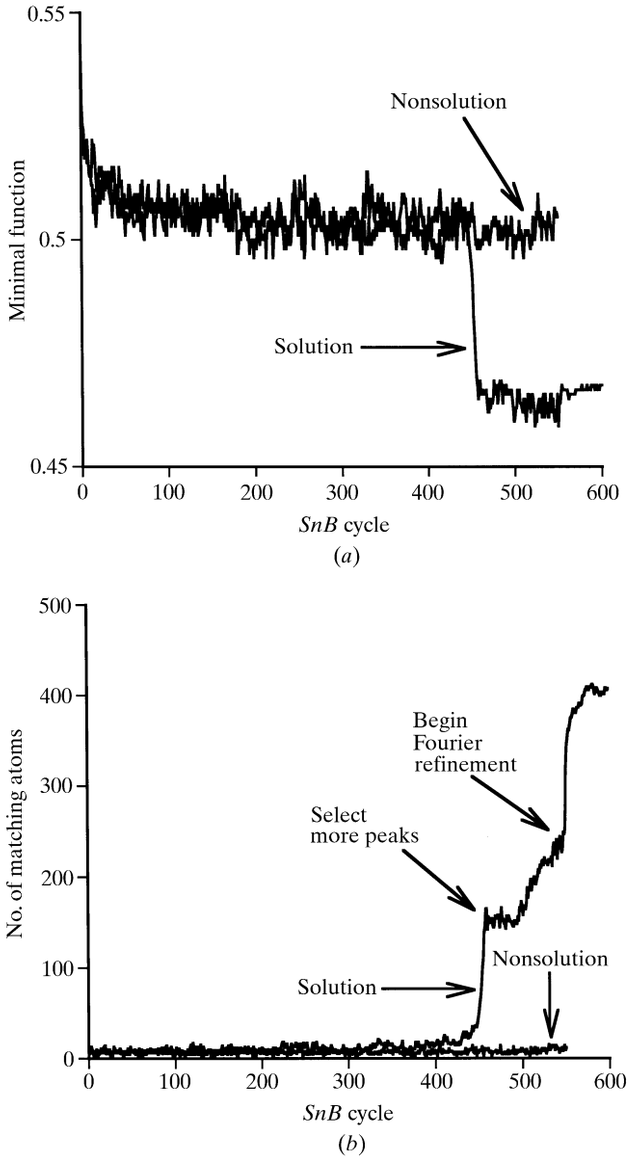

Figure 16.1.9.3
Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function *of Shake-and-Bake* cycle. (*a*) Minimal-function figure of merit, and (*b*) number of peaks closer than 0.5 Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 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 $R_{\rm unc}$ and the value of the minimal function computed using the *constrained* phases resulting from the atomic model is denoted by $R_{\rm con}$, then a function defined by

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

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.

Table 16.1.10.1
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). The largest substructures solved by these two programs are mentioned in the text.

Compound	Space group	N_u (molecule)	N_u + solvent	N_u (heavy)	Resolution (Å)	Program	Reference
Hirustasin	$P4_32_12$	402	467	10S	1.2–1.55	D	[1]
Cyclodextrin derivative	$P2_1$	448	467	—	0.88	D	[2]
Alpha-1 peptide	$P1$	408	471	Cl	0.92	S	[3]
Rubredoxin*	$P2_1$	395	497	Fe, 6S	1.0–1.1	S, D	[4]
Vancomycin	$P1$	404	547	12Cl	0.97	S	[5]
BPTI*	$P2_12_12_1$	453	561	7S	1.08	D	[6]
Cyclodextrin derivative	$P2_1$	504	562	28S	1.00	D	[7]
Balhimycin*	$P2_1$	408	598	8Cl	0.96	D	[8]
Mg-complex*	$P1$	576	608	8Mg	0.87	D	[9]
Scorpion toxin II*	$P2_12_12_1$	508	624	8S	0.96–1.2	S	[10]
Bucandin	$C2$	516	634	10S	1.05	D	[11]
Decaplanin	$P2_1$	448	635	4Cl	1.00	D	[12]
Amylose-CA26	$P1$	624	771	—	1.10	D	[13]
Viscotoxin B2	$P2_12_12_1$	722	818	12S	1.05	D	[14]
Mersacidin	$P3_2$	750	826	24S	1.04	D	[15]
Cv HiPIP H42Q*	$P2_12_12_1$	631	837	4Fe	0.93	D	[16]
Feglymycin	$P6_5$	828	1026	—	1.10	D	[17]
Acutohaemolysin	$C2_1$	1010	1242	17S	0.8	S	[18]
Tsuchimycin	$P1$	1069	1293	24Ca	1.00	D	[19]
HEW lysozyme*	$P1$	1001	1295	10S	0.85	S, D	[20], [21]
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20	D	[16]
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20	D	[22]

References: [1] Usón *et al.* (1999); [2] Aree *et al.* (1999); [3] Privé *et al.* (1999); [4] Dauter *et al.* (1992); [5] Loll *et al.* (1998); [6] Schneider (1998); [7] Reibenspiess *et al.* (2000); [8] Schäfer *et al.* (1998); [9] Teichert (1998); [10] Smith *et al.* (1997); [11] Kuhn *et al.* (2000); [12] Lehmann *et al.* (2003); [13] Gessler *et al.* (1999); [14] Pal *et al.* (2008); [15] Schneider *et al.* (2000); [16] Parisini *et al.* (1999); [17] Bunkóczi *et al.* (2005); [18] Liu *et al.* (2003); [19] Bunkóczi (2004); [20] Deacon *et al.* (1998); [21] Walsh *et al.* (1998); [22] Frazão *et al.* (1999).

It should be noted that *SHELXD* optionally deletes the highest peak if the second peak is less than a specified fraction (*e.g.* 40%) of the height of the first, in an attempt to 'kick' the structure out of a false minimum.

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 (Fig. 16.1.10.1), and there is a surprisingly high *success rate* (percentage of trial structures that go to solutions) over a narrow range of angles centred 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 would be the following. 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. Balhimycin (Schäfer *et al.*, 1998) is an example of a large non-*P*1 structure that also requires a parameter shift of around 154° to obtain a solution using the minimal function.

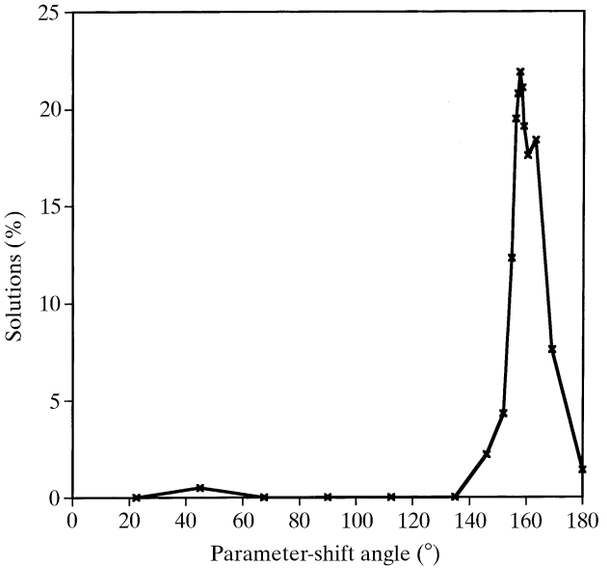

Figure 16.1.10.1
Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

16.1.10.2. Choosing a refinement strategy

Variations in the computational details of the dual-space loop can make major differences in the efficacy of *SnB* and *SHELXD*. Fig. 16.1.10.2 shows the results of different strategies tested on a 148-atom *P*1 structure (Karle *et al.*, 1989) while developing *SHELXD*. 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. Surprising, the two strategies involving random omit maps (PS + RO and TE + RO), which had been included in the test as placebos, are much more effective than the other algorithms, especially in terms of CPU efficiency. Indeed these two runs appear to approach a 100% success rate as the number of cycles becomes large. The combination of random omit maps and Karle-type tangent