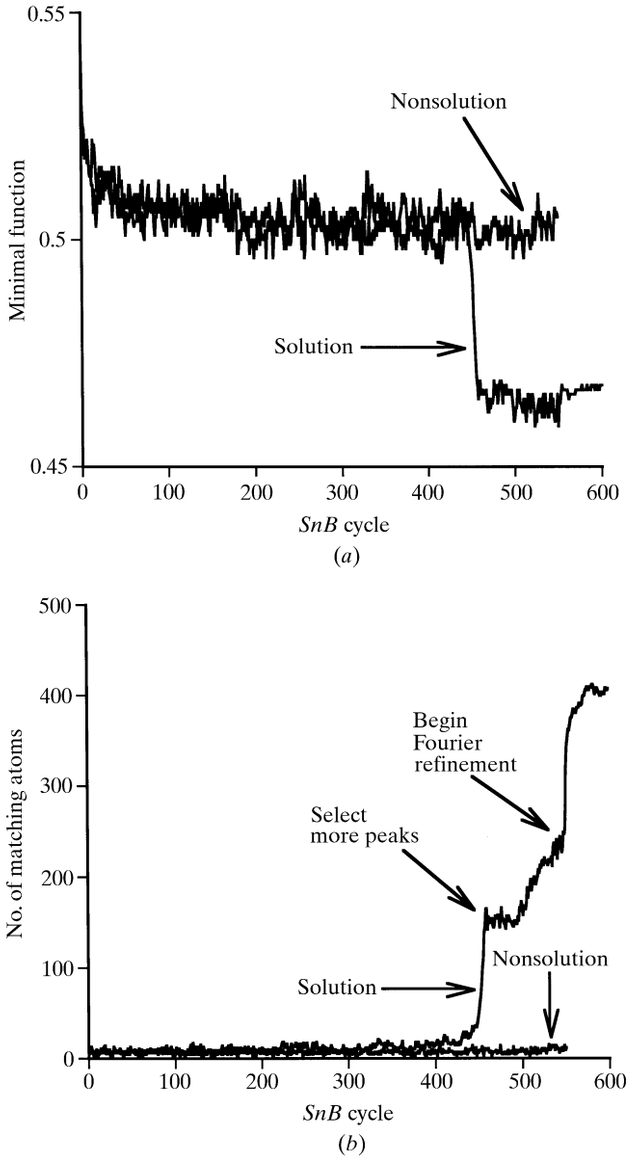


Figure 16.1.9.3
Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function *of Shake-and-Bake* cycle. (*a*) Minimal-function figure of merit, and (*b*) number of peaks closer than 0.5 Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 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.

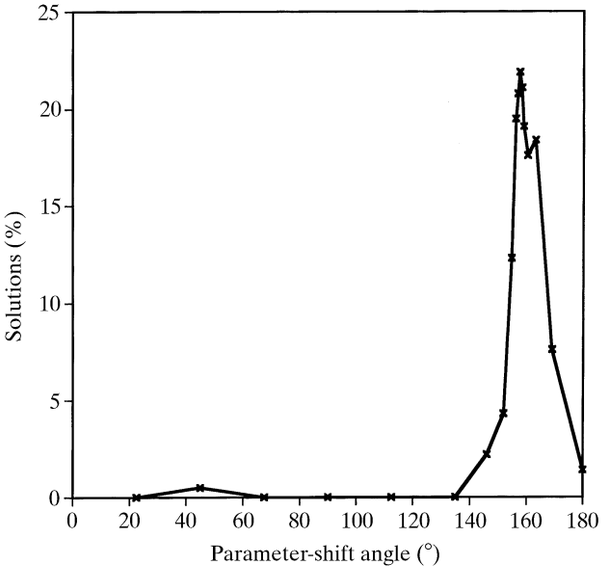
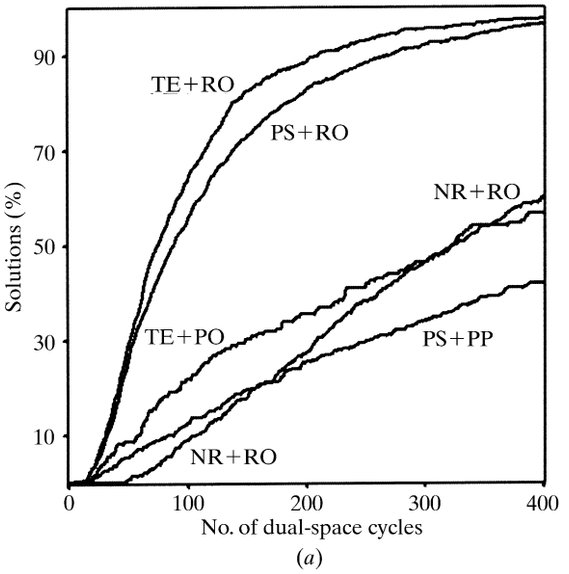


Figure 16.1.10.1
Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

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



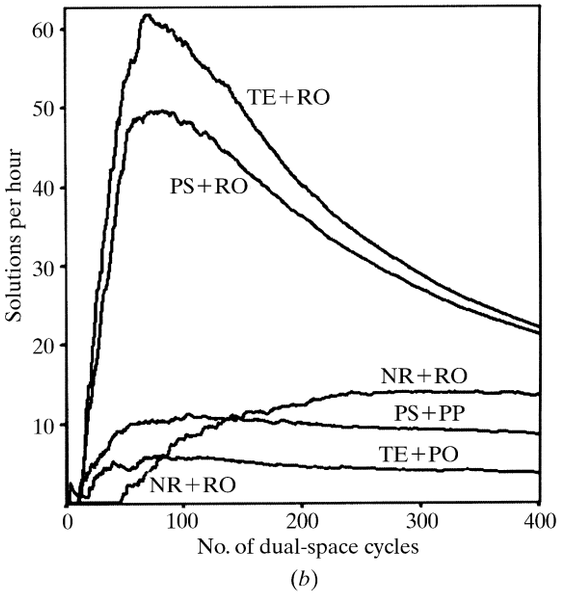


Figure 16.1.10.2
(*a*) Success rates and (*b*) cost effectiveness for several dual-space strategies as applied to a 148-atom *P*1 structure. The *phase-refinement strategies* are: (PS) parameter-shift reduction of the minimal-function value, (TE) Karle-type tangent expansion (holding the top 40% highest 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). The *real-space strategies* are: (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.

expansion appears to be even more effective (Fig. 16.1.10.3) for gramicidin A, a $P2_12_12_1$ structure (Langs, 1988). It should be noted that conventional direct methods incorporating the tangent formula tend to perform better for this space group than in *P*1, perhaps because there is less risk of a uranium-atom pseudo-solution.

Subsequent tests using *SHELXD* on several other 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

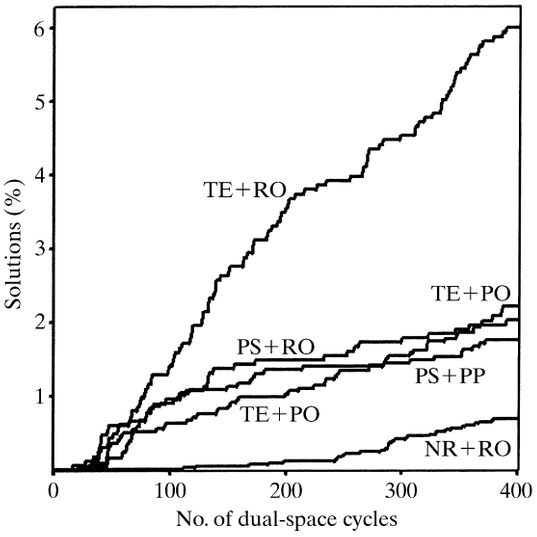


Figure 16.1.10.3
Success rates for the 317-atom $P2_12_12_1$ structure of gramicidin A.

peaklist optimization (PO) is competitive in such cases because the CPU time penalty associated with it is much smaller than when many atoms are involved.

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 the current electron density well, to perform some refinement or simulated annealing (Hodel *et al.*, 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. It is also interesting to note that the results presented in Figs. 16.1.10.2 and 16.1.10.3 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 (curves NR + RO).

16.1.10.3. Expansion to P1

The results shown in Table 16.1.1.1 and Fig. 16.1.10.2 indicate that success rates in space group *P*1 can be anomalously high. This suggests that it might be advantageous to expand all structures to *P*1 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 $P2_1$, *C*2 or $P\bar{1}$ (Chang *et al.*, 1997). Expansion to *P*1 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 *P*1 the translational search usually required for molecular replacement is not needed. Various Patterson superposition minimum functions (Sheldrick & Gould, 1995; Pavelčík, 1994) can also provide an excellent start for phase determination for data expanded to *P*1. Drendel *et al.* (1995) were successful in solving small organic structures *ab initio* by a Fourier recycling method using perturbed Fourier amplitudes and data expanded to *P*1 without the use of

probability theory. The random-omit procedure combined with expansion to *P*1 in *SHELXD* also enables structures to be solved efficiently even when the tangent formula phase extension is switched off; this has the advantage that lower E values can be used than would be suitable for the tangent formula, but at the cost of increasing the CPU time per solution. The program *ACORN2* (Dodson & Woolfson, 2009) is also particularly effective in *P*1; it applies sophisticated density modification and dual-space recycling with a special density disturbance term (POWDM) that is applied every tenth cycle. For the *P*1 form of lysozyme (see Table 16.1.10.1), good phases can be obtained by *ACORN2* starting from a fragment as small as two sulfur atoms for the 0.93 Å data. Expansion of the data to *P*1 is an essential feature of the charge-flipping approach described in Section 16.1.12.6. In general, one can say that dual-space recycling of data expanded to *P*1 requires some reasonable perturbation of the density (*e.g.* charge flipping or random peak omit) to prevent stagnation, but with this precaution provides a simple and effective approach to structure solution.

16.1.10.4. Substructure applications

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 has improved (Hendrickson & Ogata, 1997; Smith, 1998), there has been an increased need to locate many selenium sites. For larger structures (*e.g.* 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 analysed increases with the square of the number of atoms. Experimentally measured difference data are an approximation to the data for the hypothetical substructure, 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.

SHELXD is frequently successfully employed with $|F_A|$ values derived from multiwavelength MAD data generated, for example, by the programs *SHELXC* (Sheldrick, 2008, 2010) or *XPREP* (Bruker AXS, Madison, WI). The decision at which resolution the data should be truncated for substructure determination is best taken on the basis of the correlation coefficients between the signed anomalous differences (Schneider & Sheldrick, 2002). On the other hand, *SnB* is normally applied separately to anomalous and dispersive differences. In many cases, both approaches lead to successful substructure solution. The real advantage of MAD data is that they provide more experimental phase information (*i.e.* better maps) and this is most important at medium to low resolution. 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). Consideration should be limited to those data pairs $(|E_1|, |E_2|)$ [*i.e.*, isomorphous pairs $(|E_{\rm nat}|, |E_{\rm der}|)$ and anomalous pairs $(|E_{+\mathbf{H}}|, |E_{-\mathbf{H}}|)$] for which

$$\min\left[|E_1|/\sigma(|E_1|), |E_2|/\sigma(|E_2|)\right] \geq x_{\rm min} \qquad (16.1.10.2)$$

and

$$\frac{\|E_1| - |E_2\|}{[\sigma^2(|E_1|) + \sigma^2(|E_2|)]^{1/2}} \geq y_{\rm min}, \qquad (16.1.10.3)$$

where typically $x_{\rm min} = 3$ and $y_{\rm min} = 1$. The final choice of maximum resolution to be used should be based on inspection of the spherical shell averages $\langle|E_\Delta|^2\rangle_s$ *versus* $\langle s\rangle$ where $s = \sin(\theta)/\lambda$. The purpose of this precaution is to avoid spuriously large $|E_\Delta|$ values for high-resolution data pairs measured with large uncertainties due to imperfect isomorphism or general fall-off of scattering intensity with increasing scattering angle. Only those $|E_\Delta|$'s for which

$$|E_\Delta|/\sigma(|E_\Delta|) \geq z_{\rm min} \qquad (16.1.10.4)$$

(typically $z_{\rm min} = 3$) should be deemed sufficiently reliable for subsequent phasing. The probability of very large difference $|E|$'s (*e.g.* >5) 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. A paper by Blessing & Smith (1999) elaborates further data-selection criteria. 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 16.1.9.1. This will typically produce the desired overdetermination.

It is rare for Se atoms to be closer to each other than 5 Å, and the application of *SnB* to AdoHcy hydolase data truncated to 4 and 5 Å 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.

16.1.11. Substructure solution for native sulfurs and halide soaks

In the past, experimental phasing usually involved either the preparation of selenomethionine derivatives or the incorporation of heavy-metal ions by soaking crystals with a low concentration of the metal salt for several hours. The first of these methods required time in the wet lab and did not work well for all expression systems; the second had a low success rate. The improved quality of modern diffraction data collected from cryo-cooled crystals makes it now possible to exploit the weak anomalous signal from the native sulfur atoms or from halide ions introduced by soaking with a high concentration of a halide (iodide or bromide) for a few seconds immediately before cryocooling the crystal (Dauter *et al.*, 2000, 2001; Usón *et al.*,