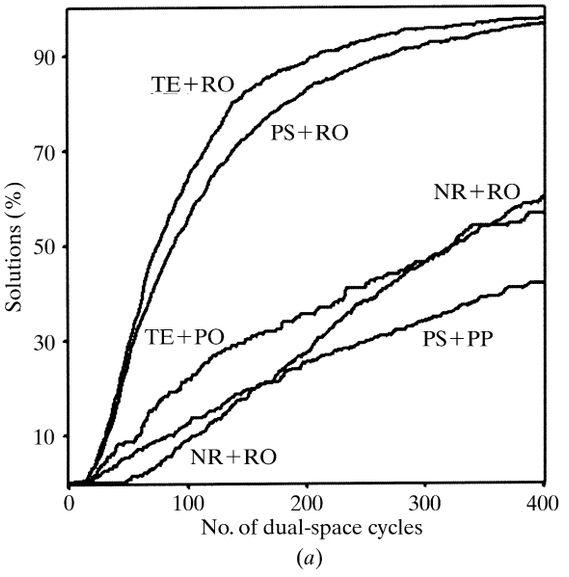

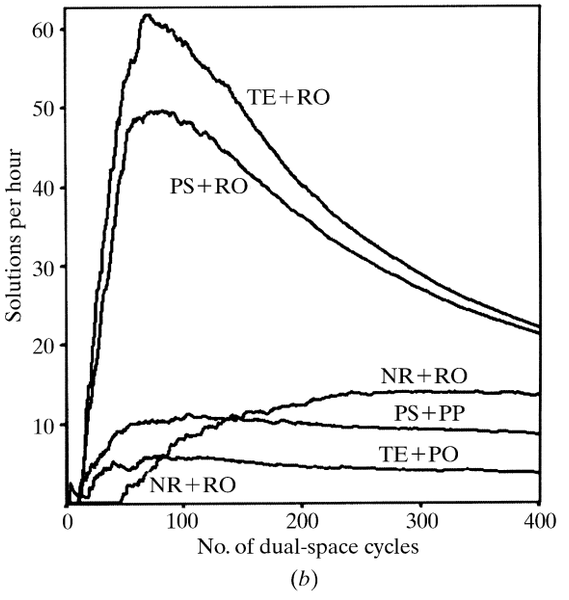

Figure 16.1.10.2
(*a*) Success rates and (*b*) cost effectiveness for several dual-space strategies as applied to a 148-atom *P*1 structure. The *phase-refinement strategies* are: (PS) parameter-shift reduction of the minimal-function value, (TE) Karle-type tangent expansion (holding the top 40% highest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

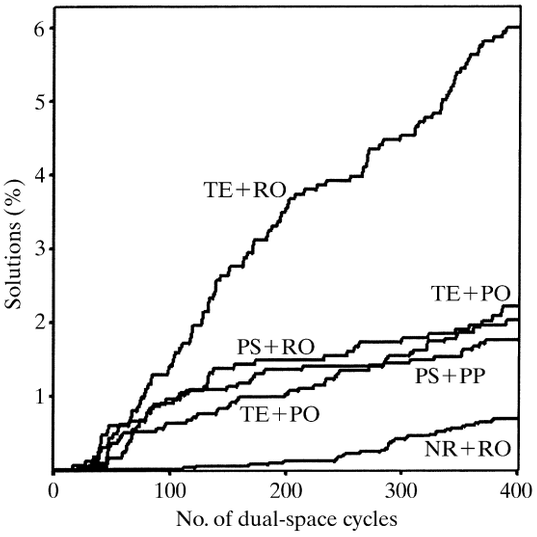

Figure 16.1.10.3
Success rates for the 317-atom $P2_12_12_1$ structure of gramicidin A.

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 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.*,