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

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

and

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

where typically $x_{\min} = 3$ and $y_{\min} = 1$. 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

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

(typically $z_{\min} = 3$) 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.*,

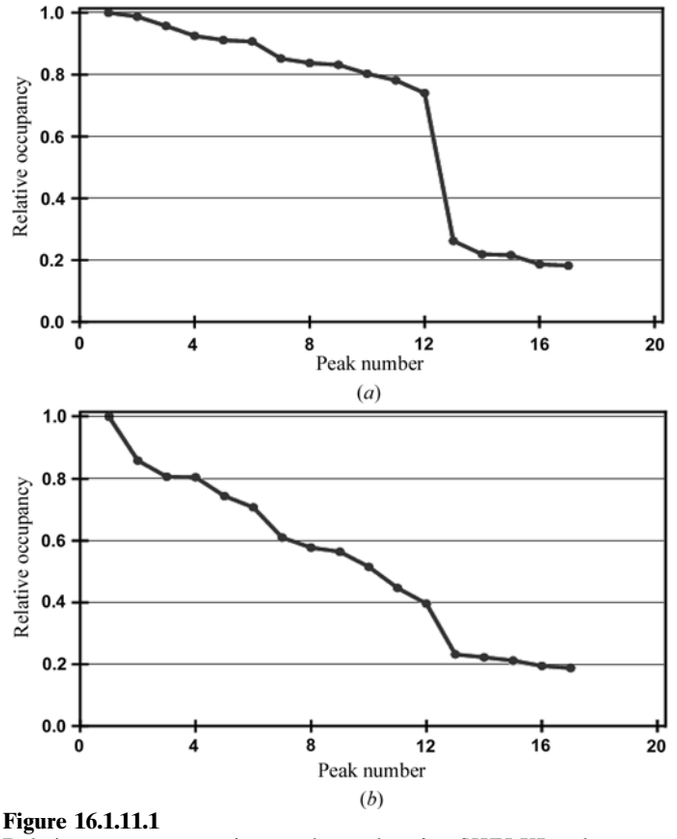

Figure 16.1.11.1
Relative occupancy against peak number for *SHELXD* substructure solutions of elastase. (*a*) Sulfur-SAD experiment showing the presence of the 12 expected sulfur atoms. (*b*) Iodide soak. Subsequent analysis showed that the peaks with relative occupancies less than 0.2 are mainly noise. These figures were made with *HKL2MAP* (Pape & Schneider, 2004).

2003). The success of these approaches is also made possible by the ability of modern, dual-space, substructure-solution programs to locate correctly a large number of sites, possibly with varying occupancies, using the SAD and SIRAS approaches.

In selenomethionine SAD and MAD phasing and in sulfur SAD phasing, the variation of the occupancies (refined in the final two cycles in the case of *SHELXD*) provides a very good indication as to whether the structure has been solved. Fig. 16.1.11.1(*a*) shows the phasing of elastase with sulfur SAD; a sharp drop in the relative occupancy after the 12th site confirms the expected presence of 12 sulfur atoms. For an iodide soak of the same protein (Fig. 16.1.11.1*b*), the relative occupancies show a gradual fall with peak number. Since the number of sites is difficult to estimate in advance for a halide soak and *SHELXD* needs to know this number approximately (within say 20%), it may be necessary to make several trials with different numbers of expected sites. From experience, the best number to use is the one that causes the occupancies to fall to about 0.2 relative to the strongest peak. Usually, subsequent refinements of the occupancies show that all the sites are partially occupied for halide soaks.

When the anomalous signal does not extend beyond about 2.0 Å, the two sulfur atoms of a disulfide bridge coalesce to a single maximum, often referred to as a supersulfur atom. At low resolution, this increases the signal-to-noise ratio for such sites in the dual-space procedure, but tends to impede phase extension to higher resolution (*e.g.* when density modification is applied to the native data with the starting phases estimated using these supersulfur atoms). An efficient way around this problem is to fit dumbbells rather than single atoms in the peak-search part of the dual-space recycling (Debreczeni *et al.*, 2003); this dramatically improves the quality of the higher-resolution starting phases.

Because the weak anomalous signal is swamped by the noise at higher resolution in such SAD experiments, it is often essential to truncate the resolution of the anomalous difference data before searching for the substructure. For MAD experiments, it is customary to truncate the data to the resolution at which the correlation coefficient between the signed anomalous differences falls below 30% (Schneider & Sheldrick, 2002). The same criterion can be used for SAD experiments if two independent data sets (*e.g.* from two different crystals) are available. As a compromise, the signed anomalous differences can be divided randomly into two sets, and then the correlation coefficient between them can be calculated. However, since these sets are not completely independent, a higher threshold (say 40%) might be advisable. An alternative criterion is to truncate the data at the point where the ratio of the mean absolute anomalous difference to its mean standard deviation falls below ~1.3, but this requires rather precise estimates of the standard deviations. In borderline cases, especially when multiple CPUs are available, it is probably safer simply to run the substructure solution for a range of different resolution cutoffs in parallel, and this is already implemented in several of the automated phasing pipelines. Sometimes good solutions are only obtained in a rather limited resolution cutoff range. A good starting value for sulfur SAD is the diffraction limit plus 0.5 Å.

16.1.12. Computer programs for dual-space phasing

Macromolecular crystallography is well served with free, high-quality, open-source software. Programs that provide direct-methods phasing for macromolecular problems will now be outlined. Although they all (except *CRUNCH2*) implement procedures that can be described more-or-less as dual-space methods, there are also appreciable differences from the three programs discussed so far. In this section, we have attempted to highlight these differences.

16.1.12.1. ACORN

ACORN (Yao *et al.*, 2006) and its successor *ACORN2* (Dodson & Woolfson, 2009) start with a fragment. This fragment can be very small: 1–8% in *ACORN*, and as little as 0.25% of the scattering is reported for *ACORN2*. Strictly speaking, these are not direct-methods programs, since they solve and refine crystal structures from poor starting phase sets that are usually derived from a known fragment. However, since this fragment can be very small, and since for *P*1 structures a single heavy atom at the origin suffices as a useable starting point, they are included here.

The data are normalized to give E magnitudes and partitioned into three sets: (1) large observed normalized magnitudes, (2) small magnitudes (typically < 0.2), and (3) the unobserved reflections (which are given values of unity) for a resolution range. A fragment is used to generate a set of phases, and this is followed by a sophisticated density-modification procedure:

$$\begin{aligned}
\rho^{(n+1)} &= 0 \quad \text{if} \quad \rho^{(n)} \leq L\sigma, \\
\rho^{(n+1)} &= \rho^{(n)} \tanh[0.2(\rho^{(n)}/\sigma)^{\eta}] \quad \text{if} \quad \rho^{(n)} > L\sigma, \\
\rho^{(n+1)} &= T\sigma \quad \text{if} \quad \rho^{(n+1)} > T\sigma,
\end{aligned} \tag{16.1.12.1}$$

where σ is the standard deviation of the map density and