International Tables for Crystallography (2012). Vol. F, Section 16.1.12.1, pp. 426–427.

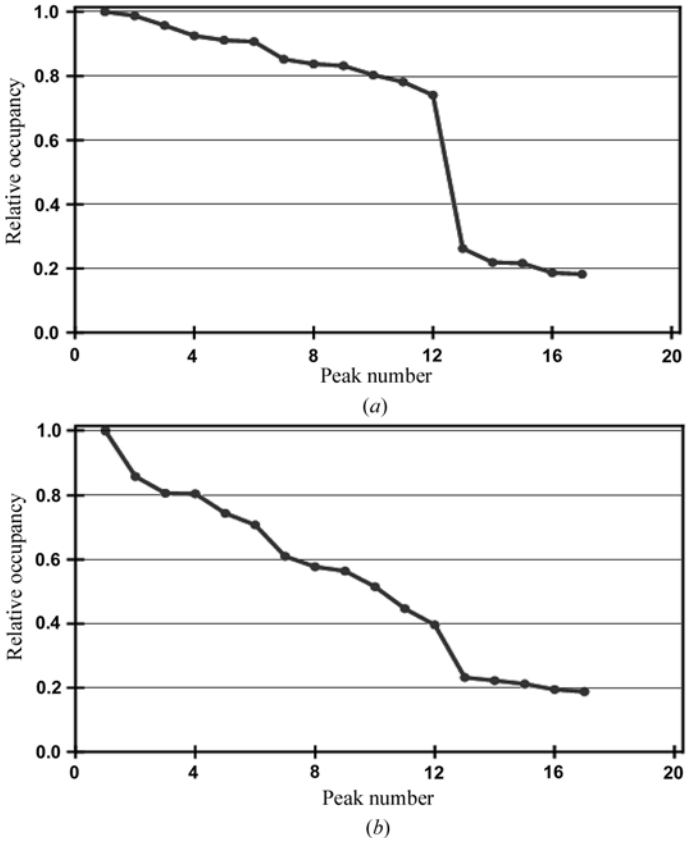

Figure 16.1.11.1
Relative occupancy against peak number for *SHELXD* substructure solutions of elastase. (*a*) Sulfur-SAD experiment showing the presence of the 12 expected sulfur atoms. (*b*) Iodide soak. Subsequent analysis showed that the peaks with relative occupancies less than 0.2 are mainly noise.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

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

where σ is the standard deviation of the map density and

$$T = \max(T_1 + c + 0.5c^2, 100),$$

$$T_1 = \left(\frac{M}{N}\right)^{1/2} \frac{Z_{\max}}{14}; \quad 3 \leq T_1 \leq 15, \qquad (16.1.12.2)$$

where M is the number of observable reflections within the resolution sphere and N is the number of atoms in the unit cell (excluding H atoms). The unconstrained value of T_1 is approximately 0.5 of the expected peak height of the heaviest atom in the E map with perfect phases; c is the cycle number.

$$L = L_1 - L_1^{c/n},$$

$$L_1 = 1.05\left[\left(B/r^2\right) - 1\right]\Phi(Z_{\max}), \qquad (16.1.12.3)$$

where B is the usual overall temperature factor; Φ is a cubic function going through the points (Φ, Z) = (0.84, 16), (0.96, 30), (1.15, 34.5) and (1.24, 48). If $Z_{\max} < 16$, the value $Z = 16$ (corresponding to sulfur) is used, and for $Z_{\max} > 48$, the value $Z = 48$ (corresponding to cadmium) is used. The value of L is thus reduced in n cycles from L_1 to zero.

$$n = \mathrm{nint}(0.5/p), \qquad (16.1.12.4)$$

where 'nint' indicates the nearest integer and

$$p = \frac{\sum_{\text{fragment}} Z^2}{\sum_{\text{all atoms}} Z^2}. \qquad (16.1.12.5)$$

Finally,

$$\eta = 17.24(r-1)^5 + 1.5. \qquad (16.1.12.6)$$

This use of η improves performance at low resolution. There is also an E limit,

$$E_{\text{lim}} = [0.15 + (1/r)](0.85 + 10p), \qquad (16.1.12.7)$$

imposed with the proviso that $0.75 \leq E_{\text{lim}} \leq 1.25$. Resolution enhancement is simple: all missing data can optionally be given a value of 1.0, which is the square root of the expectation value of E^2 from Wilson statistics. It can be seen that this procedure always needs a fragment, but this can be very small indeed. Examples include five S atoms in the catalytic domain of chitinase A1 from *Bacillus circulans* WL-12 (Matsumoto *et al.*, 1999) that account for less than 1% of the scattering density and oxidoreductase (Haynes *et al.*, 1994) with less than 0.5% of the scattering power in the initial fragment. In the latter case, the space group is $P1$ so one heavy atom can be placed at the unit-cell origin. Two atoms define the fragment: one was placed at the origin and the second in a position compatible with the Patterson map. The final mean phase error was 31.6°. *ACORN2* is available as part of the *CCP4* package at http://www.ccp4.ac.uk.

16.1.12.2. *IL MILIONE*

The *IL MILIONE* software is a product of the Bari group (Burla *et al.*, 2007). It provides a complete suite of programs for structure solution including software for processing SAD, MAD or SIRAS data or for molecular replacement. We will focus here on the *ab initio* procedures in this package. *Ab initio* phasing uses triplets and tangent procedures or Patterson methods. Direct-space refinement using density modification is employed along with the use of a resolution extension procedure. For the initial phasing, triplet invariants are evaluated by means of the P10 formula (Cascarano *et al.*, 1984). The tangent formula [equation (16.1.4.1)] is used in conjunction with these triplet-phase estimates starting with random phases and multiple starting points. An early figure of merit (eFOM) is calculated for each tangent trial and only the best trial solutions based on this are submitted to direct-space refinement. When misplaced molecular fragments are present, the structures can often be solved by the *RELAX* procedure (Burla *et al.*, 2003). In this procedure, the phases of a trial solution obtained in the correct space group are extended and refined in $P1$ by direct-space techniques. The appropriate figures of merit are used to determine the appropriate vector shift to operate on the fragment.

Patterson deconvolution may also be used in conjunction with direct methods. A superposition minimum function is used. The first peak from this procedure is always used in the phasing process. For each of the remaining peaks (the number of which depends on the size of the structure to be solved and on its data resolution), a set of phases is obtained which is ranked by a specific early figure of merit (pFOM) defined as

$$\text{pFOM} = \sigma/R(\Phi), \qquad (16.1.12.8)$$

where σ is the usual standard deviation of the electron-density map and $R(\Phi)$ is the minimal function [equation (16.1.4.2)]. Irrespective of whether direct or Patterson methods are used, direct-space refinement techniques consist of cycles of electron-density modification (in which only a small fraction of the electron densities are inverted to obtain new phases), and a mask from molecular envelope calculations is applied. The correct solution is identified by a final figure of merit (fFOM), and the program automatically stops when fFOM exceeds a given threshold that depends on data resolution and structural complexity. The Patterson deconvolution methods proved to be, by far, the most efficient ones for large structures and, therefore, have been chosen as the default procedure in this case. They extended the size of the macromolecular structures that are solvable *ab initio* to more than 6000 non-H atoms in the asymmetric unit, provided that at least one calcium atom is present in the asymmetric unit and atomic resolution data are available (Burla *et al.*, 2007).

In favourable circumstances, *IL MILIONE* is also able to solve protein structures with data resolution up to 1.4–1.5 Å and to provide interpretable electron-density maps. The program has been tested using about 100 structures randomly taken from among those in the PDB with resolution better than 1.6 Å. It was able to solve all the test structures that had atomic resolution data, less than 2000 non-H atoms in the asymmetric unit (Nasym), and atoms heavier than calcium present (Zmax > 20). The solution efficiency is reduced to 84% for structures with atomic resolution data, Zmax > 20 and Nasym > 2000, and it is reduced further when these conditions are not fully met. In the presence of atoms as heavy as Ho, Au, Hg or Yb, solutions of structures composed of more than 1000 atoms have been achieved at resolutions as low as 2.0 Å. Finally, it was successful even at 1.65 Å for a case containing up to 7890 non-hydrogen atoms in the asymmetric unit (Caliandro *et al.*, 2008).

IL MILIONE can also apply direct methods to SIR and MIR data. Two different approaches may be followed for protein crystal-structure solution from isomorphous data (up to five derivatives may be used). The triplet phase invariants are estimated *via* the conditional probability distribution function,

$$P(\Phi_p|E_{p\mathbf{H}}, E_{p\mathbf{K}}, E_{p\mathbf{H+K}}, E_{d\mathbf{H}}, E_{d\mathbf{K}}, E_{d\mathbf{H+K}}) = [2\pi I_0(G\cos\Phi_p)] \qquad (16.1.12.9)$$

(Hauptman, 1982*a*; Giacovazzo *et al.*, 1988, 1996), where Φ_p is the triplet phase of the protein and

references