International Tables for Crystallography (2012). Vol. F, Section 16.1.12.2, pp. 427–428.

$$T = \max(T_1 + c + 0.5c^2, 100),$$
$$T_1 = \left(\frac{M}{N}\right)^{1/2} \frac{Z_{\max}}{14}; \quad 3 \le T_1 \le 15, \qquad (16.1.12.2)$$

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.

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

where 'nint' indicates the nearest integer and

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

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,

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

imposed with the proviso that $0.75 \le E_{\rm lim} \le 1.25$. 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

$$G = 2(\sigma^3/\sigma_2^{3/2})_p E_{p\mathbf{H}} E_{p\mathbf{K}} E_{p\mathbf{H+K}} + 2q(\sigma^3/\sigma_2^{3/2})_{\mathbf{H}} \Delta_{\mathbf{H}} \Delta_{\mathbf{K}} \Delta_{\mathbf{H+K}}$$
$$\Delta = (F_d - F_p)/\Sigma_H^{1/2}. \qquad (16.1.12.10)$$

The factor q takes into account lack of isomorphism and measurement errors, and the Δ parameters are isomorphous differences normalized with respect to the heavy-atom structure. Φ_p is expected to be close to 0 or π according to whether G is positive or negative. A starting set of phases is generated by a random process, a weighted tangent formula is applied to these phases and various trials are produced among which the correct solution may be found by a suitable figure of merit. If multiple derivatives are available, the program is able to estimate automatically, for each derivative, the scattering power of the heavy-atom structure and also to suggest which is the best derivative. The *IL MILIONE* package can be obtained from http://www.ic.cnr.it.

16.1.12.3. SHELX

The *SHELX* family of programs is widely used for small- to medium-sized structure solution and refinement. The family also contains three programs that are extensively used in macromolecular crystallography: *SHELXC*, *SHELXD* and *SHELXE*. For an overview of the *SHELX* system, see Sheldrick (2008). *SHELXC* is a housekeeping program designed to prepare the necessary files for *SHELXD* and *SHELXE*. *SHELXD* (Sheldrick, 1998; Schneider & Sheldrick, 2002; Usón & Sheldrick, 1999) is employed both for substructure solution and for *ab initio* direct methods for atomic resolution data as described elsewhere in this chapter. Fu *et al.* (2007) have shown how to adapt it to multiple CPU systems. *SHELXE* (Sheldrick, 2002, 2008, 2010) improves experimental phases from SAD, SIRAS or MAD data or starting phases from molecular replacement by iterative density modification and autotracing. *SHELX* can be obtained at http://shelx.uni-ac.gwdg.de/SHELX/.

16.1.12.4. SnB and BnP

SnB was the first program to solve small macromolecules *ab initio*, using a global cost function [equation (16.1.4.2)] that reflects how well the calculated phases fit the expected distribution of the triplets. It is fully described elsewhere in this chapter and is an effective tool in structure and substructure determination. Versions are available for multiple CPU systems (Rappleye *et al.*, 2002) and computational grids (Miller *et al.*, 2007). It is also available as part of the *BnP* package (Weeks *et al.*, 2002) that was produced in collaboration with the Biocrystallography Laboratory at the University of Pittsburgh for the experimental phasing of macromolecules. *SnB* is available from http://www.hwi.buffalo.edu/SnB/ and *BnP* from http://www.hwi.buffalo.edu/BnP/.

16.1.12.5. HySS

The substructure solution program *HySS* (Grosse-Kunstleve & Adams, 2003), which is part of the *PHENIX* package (Adams *et al.*, 2007), is closely modelled on *SHELXD* but was implemented using the cctbx libraries (Grosse-Kunstleve *et al.*, 2002). The main differences to *SHELXD* are (1) the translational search for two-atom fragments is performed by Fourier methods followed by a peak search rather than a random search, (2) the use of the tangent formula in reciprocal space is replaced by squaring the density in real space, and (3) several termination criteria are implemented so that the program can stop when the structure appears to be solved. The *PHENIX* package can be obtained at http://www.phenix-online.org/.

16.1.12.6. SUPERFLIP: charge flipping

Charge flipping is a disturbingly simple dual-space algorithm (Oszlányi & Sütő, 2004, 2005, 2008). It uses as input only the cell parameters of the structure, the reflection indices and the intensities. The intensities can be corrected for thermal motion *via* an overall temperature factor if required, and this is often beneficial. Neither chemical information nor the symmetry is explicitly used in the structure solution process. The electron density is sampled on a discrete rectangular grid of pixels with values ρ_i, $i = 1, N_{\rm pix}$. The algorithm proceeds iteratively. To begin the process, a starting set of structure factors is created by combining the experimental structure-factor amplitudes with random phases. Each iteration or cycle (numbered n) involves four steps:

(1) A trial electron density $\rho^{(n)}$ is obtained by inverse Fourier transform of the structure factors in the usual way.

(2) A modified density $g^{(n)}$ is obtained from $\rho^{(n)}$ by reversing the sign (charge flipping) of all density pixels with density below a certain positive threshold δ as follows:

$$\begin{aligned} g_i^{(n)} &= \rho_i^{(n)} \quad \text{if} \quad \rho_i^{(n)} > \delta, \\ g_i^{(n)} &= -\rho_i^{(n)} \quad \text{if} \quad \rho_i^{(n)} \le \delta. \end{aligned} \qquad (16.1.12.11)$$

(3) Modified structure factors are obtained by Fourier transform of $g^{(n)}$,

$$G_{\mathbf{H}}^{(n)} = FT[g^{(n)}]. \qquad (16.1.12.12)$$

(4) The structure factors $F_{\mathbf{H}}^{(n+1)}$ are obtained from $F_{\mathbf{H}}^{(n+1)}$ and $G_{\mathbf{H}}^{(n+1)} = |G_{\mathbf{H}}^{(n+1)}| \exp(2\pi i \varphi_{\mathbf{H}}^G)$ as follows:

(*a*) $F_{\mathbf{H}}^{(n+1)} = |F_{\mathbf{H}}| \exp(2\pi i \varphi_{\mathbf{H}}^G)$ for $|F_{\mathbf{H}}|$ observed and strong.

(*b*) $F_{\mathbf{H}}^{(n+1)} = |G_{\mathbf{H}}^{(n)}| \exp[2\pi i(\varphi_{\mathbf{H}}^G + 0.25)]$ for $|F_{\mathbf{H}}|$ observed and weak. In other words, for these reflections, calculated moduli are accepted unchanged and calculated phases are shifted by a constant $\Delta\varphi = \pi/2$. This means that the observed data of weak reflections are not used actively in the process, except for the knowledge that they are indeed weak. Experience shows that about 20–40% of all reflections can be considered weak. This use of the phase shifting of the weak reflections significantly improves the performance of the algorithm in cases of more complex structures (Oszlányi & Sütő, 2005); in some cases the success rate is increased by a factor of ten, in other cases a previously unsolvable structure becomes solvable by the modified algorithm.

(*c*) $F_{\mathbf{H}}^{(n+1)} = 0$ for $|F_{\mathbf{H}}|$ unobserved.

(*d*) $F_{\mathbf{H}}^{(n+1)} = G_{\mathbf{H}}^{(n)}$ for $\mathbf{H} = 0$. In other words, the value of F_{000} is not fixed.

The new set of structure factors enters the next cycle or iteration. The cycles are repeated until the calculation converges. Progress is monitored by a conventional R factor where a small change in R signals convergence. The parameter δ is the main variable of the iteration, and its value can be critical. It must often be determined by trial and error, but this search can be automated. The second variable parameter of the algorithm is the proportion of the reflections considered weak in each cycle.