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

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:

$$g_i^{(n)} = \rho_i^{(n)} \quad \text{if} \quad \rho_i^{(n)} > \delta,$$
$$g_i^{(n)} = -\rho_i^{(n)} \quad \text{if} \quad \rho_i^{(n)} \leq \delta. \quad (16.1.12.11)$$

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

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

(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.

The algorithm seeks a Fourier map that is stable against repeated flipping of all density regions below δ. Obviously, a large number of missing reflections will make the algorithm less efficient, because the missing reflections are assigned a zero amplitude, which induces large termination ripples in the Fourier map. The underlying assumption of the algorithm, that the density is close to zero in large regions of the unit cell and positive in small parts of the unit cell, is no longer fulfilled and the algorithm fails. The question of incomplete data has been addressed by Palatinus *et al.* (2007). They show that the missing data can be approximated on the basis of the Patterson map of the unknown structure optimized by the maximum-entropy method. Structures that could not be solved by the original charge-flipping algorithm can be solved in this way. For small molecules, 50% or more of the reflections can be missing, and the structure can still be reconstructed by charge flipping. The situation for macromolecules is less clear.

Symmetry is an important issue. Surprisingly, in the charge-flipping method, all structures are solved in space group $P1$, and all symmetry constraints are ignored. Attempts to impose symmetry usually damage the process fatally. The disadvantage of this is that the charge density of the whole unit cell must be determined, and not just that of the asymmetric unit. Furthermore, the symmetry elements must be located once a solution has been found. A computer program, *SUPERFLIP* (Palatinus & Chapuis, 2007), and a Java applet that demonstrates the procedure in two dimensions are freely available for download at http://escher.epfl.ch/flip/.

The charge-flipping method has been adapted to proteins (Dumas & van der Lee, 2008) and applied to a $P1$ structure with 7111 atoms [*i.e.* liver alcohol dehydogenase in complex with NADH and Cd-DMSO: 5866 protein atoms, 1241 waters and 4 Cd atoms (Meijers *et al.*, 2007)]. In common with other methods described in this chapter, charge flipping is much more effective for data to very high resolution (in this case 1.0 Å) and especially for structures containing heavier atoms. The method can also, in principle, be used for substructure determination; the solution of known substructures with as many as 120 unique Se atoms is reported in the same paper.

16.1.12.7. CRUNCH2 – Karle–Hauptman determinants

The program *CRUNCH2* is quite different to the other programs mentioned in this section. With the exception of some *E*-map recycling at the end to complete a substructure, *CRUNCH2* operates entirely in reciprocal space by maximizing higher-order Karle–Hauptman determinants (Karle & Hauptman, 1950; de Gelder *et al.*, 1993). It is incorporated into the automated *CRANK* pipeline for macromolecular structure solution (Ness *et al.*, 2004). The quality of the substructure solutions obtained appears to be at least as good as those from the dual-space programs, but it may be slower for large substructures.

16.1.13. Conclusions and the grand challenge

In practice, the main use of direct methods in macromolecular crystallography is to obtain substructures using SAD and MAD data where the limitations of the method can be relaxed. There are, of course, a few structures solved *ab initio*, but they are relatively uncommon. There is a grand challenge here: to solve *ab initio* macromolecular structures using the native data alone at resolutions more typical for macromolecules without the need for specific prior structural knowledge.

The extensive (and successful) use of atomicity constraints both in real space (peak picking) and reciprocal space (tangent formula and minimal function) make it difficult to overcome the need for atomic resolution data in the *Shake-and-Bake* methods. At lower resolution, the atomicity constraint should be replaced by another based on the recurrence of model fragments that can be predicted *a priori* from the protein sequence (*e.g.* small polyalanine α-helices, β-sheets, cofactors, bases, disulfide bridges *etc.*). The effectiveness of a very small, yet accurate, fraction of the total scattering mass in the form of a fragment or heavy atoms is apparent from the results of *ACORN2* and *IL MILIONE*.

Shortly before this chapter went to press, a paper by the Usón group (Rodríguez *et al.*, 2009) showed a possible way ahead in the case of equal-atom structures, by exploiting general features of protein secondary structure. In its current form, the method requires that the protein is at least 20% α-helical and diffracts to 2.0 Å or better, requirements that would be fulfilled by at least a quarter of the protein crystal structures deposited in the PDB. The method was successfully applied to four test structures and one previously unsolved 222 amino-acid structure that diffracted to 1.95 Å and had resisted all previous attempts at solution by molecular replacement and experimental phasing. The method exploits the power of the molecular-replacement program *PHASER* (McCoy *et al.*, 2007) to search for multiple copies of (for example) 14-residue α-helices with data truncated to 2.5 Å, retaining several thousand 'best' solutions at each stage as judged by maximum-likelihood criteria. These potential multi-helix solutions are all input into a new version of the program *SHELXE* (Sheldrick, 2010) that applies density modification and main-chain tracing iteratively. At some point, depending on the size of the structure and the quality of the data, but typically for a trial structure consisting of three or four α-helices making up some 12% of the structure, the autotracing locks in and gives a relatively complete backbone trace that can be immediately recognized both by the number of connected residues traced and a correlation coefficient between the calculated and observed E values. A multiple CPU computer grid is essential for performing these numerically intensive calculations in parallel, and the whole branching and pruning operation is performed under the control of the program *ARCIMBOLDO*. This approach is still at an early stage and should benefit from fine-tuning and the inevitable future increases in computer power, but it clearly has the potential to become a main-stream *ab initio* method for the solution of protein structures.

The development, in Buffalo, of the *Shake-and-Bake* algorithm and the *SnB* program has been supported by grants GM-46733 from NIH and ACI-9721373 from NSF, and computing time from the Center for Computational Research at SUNY Buffalo. HAH, CMW and RM would also like to thank the following individuals: Chun-Shi Chang, Ashley Deacon, George DeTitta, Adam Fass, Steve Gallo, Hanif Khalak, Andrew Palumbo, Jan Pevzner, Thomas Tang and Hongliang Xu, who have aided the development of *SnB*, and Steve Ealick, P. Lynne Howell, Patrick Loll, Jennifer Martin and Gil Privé, who have generously supplied data sets. The development, in Göttingen, of *SHELXD* has been supported by the BIOXHIT Consortium and the HCM Institutional Grant ERB CHBG CT 940731 from the European Commission. GMS and IU wish to thank Thammarat Aree, Gábor Bunkóczi, Zbigniew Dauter, Judit É. Debreczeni, Judith