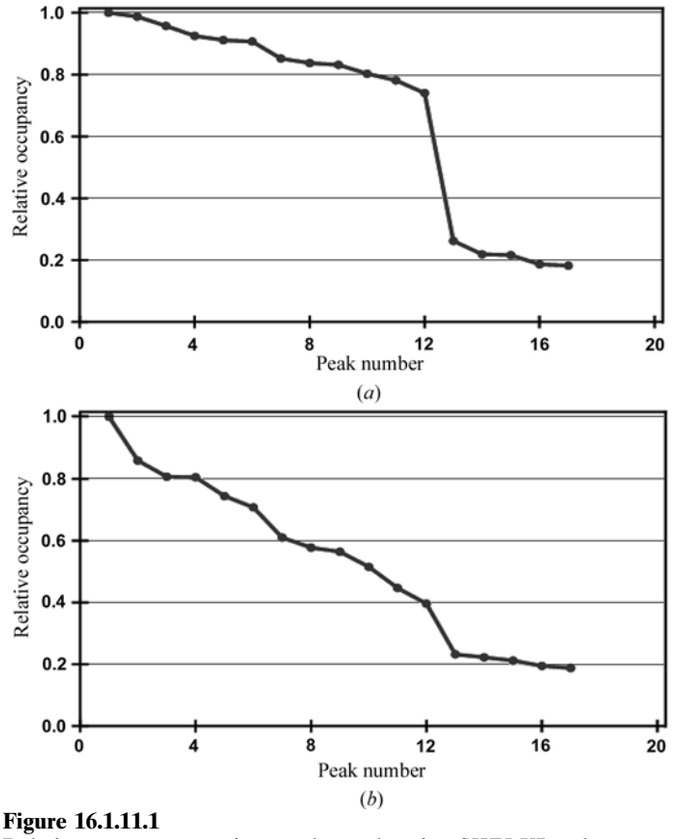


**Figure 16.1.11.1**
Relative occupancy against peak number for *SHELXD* substructure solutions of elastase. (*a*) Sulfur-SAD experiment showing the presence of the 12 expected sulfur atoms. (*b*) Iodide soak. Subsequent analysis showed that the peaks with relative occupancies less than 0.2 are mainly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

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

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; $\Phi$ is a cubic function going through the points ($\Phi$, $Z$) = (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.

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

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 $\eta$ 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 $0.75 \le E_{\text{lim}} \le 1.25$. 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 $\sigma$ 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,

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

(Hauptman, 1982*a*; Giacovazzo *et al.*, 1988, 1996), where $\Phi_p$ 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 $\Delta$ parameters are isomorphous differences normalized with respect to the heavy-atom structure. $\Phi_p$ is expected to be close to 0 or $\pi$ 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 $\rho_i$, $i = 1, N_{\rm pix}$. 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 $\delta$ 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 $\delta$ 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 $\delta$. 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 $\alpha$-helices, $\beta$-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% $\alpha$-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 $\alpha$-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 $\alpha$-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