A special version of *SHELXD* is being developed that makes extensive use of the Patterson function both in generating starting atoms and in providing an independent figure of merit. It has already successfully located the anomalous scatterers in a number of structures using MAD F_A data or simple anomalous differences. A recent example was the unexpected location of 17 anomalous scatterers (sulfur atoms and chloride ions) from the 1.5 Å-wavelength anomalous differences of tetragonal HEW lysozyme (Dauter *et al.*, 1999).

16.1.9. Extending the power of direct methods

The *Shake-and-Bake* approach has increased, by an order of magnitude, the size of structures solvable by direct methods. In addition, a routine application of the *SnB* program to peak-wavelength anomalous difference data has revealed 64 of the 70 Se sites in a selenomethionine-substituted protein (Deacon & Ealick, 1999). Although there is no indication that maximum size limitations have been reached, the fact that the reliability of invariant estimates is known to decrease with increasing structure size suggests that such limitations may exist; based on preliminary tests, it is conjectured that the limit is a few thousand unique atoms for conventional full-structure experiments. Thus, it is natural to wonder what can be done in situations where direct methods are not now routinely applicable. These cases include (1) macromolecules that lack heavy-atom or anomalous-scattering sites with sufficient phasing power for present techniques, (2) macromolecules for which no derivatives are available or for which selenium substitution is impossible, and (3) structures of any size which fail to diffract at sufficiently high resolution. 'Sufficiently high' typically means about 1.2 Å in non-substructure situations.

The requirement for data to very high resolution is, of course, troublesome for macromolecules. One approach to lowering resolution requirements might be to replace the peak search by a search for small common fragments (*e.g.* the five atoms of a peptide unit or an aromatic residue). Furthermore, it should also be possible to integrate the *wARP* procedure (Lamzin & Wilson, 1993; Perrakis *et al.*, 1997) into the real-space part of the *Shake-and-Bake* cycle. The Patterson function (Pavelčík, 1994; Sheldrick & Gould, 1995) and large Karle–Hauptman determinants (Vermin & de Graaff, 1978) might also improve the success rate in borderline cases by providing better-than-random starting coordinates or phases.

However, it is not necessarily true that peak picking is the primary limitation to lower-resolution applications. The lack of enough sufficiently accurate triplet-invariant values appears to be a more fundamental problem. Simulation experiments have shown that the *SnB* program can solve the crambin structure even at 2.0 Å if the invariants used are accurate enough (Weeks *et al.*, 1998). Therefore, the primary breakdown of *Shake-and-Bake* occurs in reciprocal space and could likely be overcome if correct individual invariant values were used instead of the rather crude estimates provided by the Cochran (1955) distribution for the cosines of the triplet invariants. Individual invariant estimates, $\omega_{\mathbf{HK}}$, can be accommodated by a *modified tangent formula*,

$$\tan\varphi_{\mathbf{H}} = \frac{\sum_{\mathbf{K}} W_{\mathbf{HK}} \sin(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} W_{\mathbf{HK}} \cos(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})}, \qquad (16.1.9.1)$$

or by a *modified minimal function*,

$$R(\Phi) = (1/2\textstyle\sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}}) \sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}} \{[\cos(\Phi_{\mathbf{HK}}) - \cos(\omega_{\mathbf{HK}})]^2 + [\sin(\Phi_{\mathbf{HK}}) - \sin(\omega_{\mathbf{HK}})]^2\}, \qquad (16.1.9.2)$$

where $W_{\mathbf{HK}}$ are appropriately chosen weights. Either of these relationships can serve as the basis for a modified *Shake-and-Bake* procedure.

One approach to providing better invariant values is to estimate them individually from the known structure-factor magnitudes ($|E|$'s). Several methods for doing this have been proposed over the years for the small-molecule case (*e.g.* Hauptman *et al.*, 1969; Langs, 1993), and this approach has met with limited success. In the macromolecular case, however, better options for estimating invariant values are available whenever supplemental information in the form of isomorphous-replacement or anomalous-dispersion data is provided. In addition, the development of multiple-beam diffraction raises the possibility of measuring invariant values experimentally. The modified tangent and minimal-function formulas provide the foundation for a unified treatment of all such supplemental information.

16.1.9.1. *Integration with isomorphous replacement*

The integration of traditional direct methods with isomorphous replacement was initiated by Hauptman (1982*a*), who studied the conditional probability distribution of triplet invariants comprised jointly of native and derivative phases assuming as known the six magnitudes associated with reciprocal-lattice vectors **H**, **K** and $-\mathbf{H}-\mathbf{K}$. It was shown that many triplets, whose true values were near either 0 or π, could be identified and reliably estimated. Later it was shown that cosine estimates could be obtained anywhere in the range -1 to $+1$ (Fortier *et al.*, 1985). In a series of six recent papers, Giacovazzo and collaborators utilized a combined direct-methods/isomorphous-replacement approach, with limited success, to devise procedures for the *ab initio* solution of the phase problem for macromolecules (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995; Giacovazzo & Platas, 1995; Giacovazzo, Siliqi & Platas, 1995; Giacovazzo *et al.*, 1996). Their methods depend only on diffraction data for a pair of isomorphous structures and do not require any prior structural knowledge. Hu & Liu (1997) have generalized the earlier work to obtain the conditional distribution of the general (n-phase) structure invariant when diffraction data are available for any number (m) of isomorphous structures. Finally, it has been shown that, provided the heavy-atom substructure is known, Hauptman's triplet distribution leads to unique values for the triplets and the individual phases (Langs *et al.*, 1995).

16.1.9.2. *Integration with anomalous dispersion*

In a manner analogous to the SIR case, Hauptman (1982*b*) derived the conditional probability distribution for triplet invariants given six magnitudes ($|E_{\mathbf{H}}|, |E_{-\mathbf{H}}|, |E_{\mathbf{K}}|, |E_{-\mathbf{K}}|, |E_{\mathbf{H}+\mathbf{K}}|, |E_{-\mathbf{H}-\mathbf{K}}|$) in the presence of anomalous dispersion. It was shown that unique estimates, lying anywhere in the whole interval 0–2π, could be obtained for the triplet values. This result was unanticipated since all earlier work had led to the conclusion that a twofold ambiguity in the value of an individual phase was intrinsic to the SAS approach. Later, it was demonstrated how the probabilistic estimates led to individual phases by means of a system of SAS tangent equations (Hauptman, 1996). Although the initial application of this tangent-based approach to the previously known macromomycin structure (750 non-H protein atoms plus 150 solvent molecules) was encouraging, it has not yet been applied to unknown macromolecules.

The conditional probability distributions of the quartet invariants, in both the SIR and SAS cases, have been derived based on corresponding difference structure factors rather than on the individual structure factors themselves (Kyriakidis *et al.*, 1996). Fan and his collaborators (Fan *et al.*, 1984; Fan & Gu, 1985; Fan *et al.*, 1990; Sha *et al.*, 1995; Zheng *et al.*, 1996) have also extensively studied the use of direct methods in the SAS case. Applications to the known small protein avian pancreatic polypeptide at 2 Å

revealed the essential features of the molecule. The direct-methods approach was used to break the phase ambiguity for core streptavidin and azurin II (proteins of moderate size) using SAS data at 3 Å. Although the direct-methods maps in these cases did not reveal the structures, the phases were good enough to serve as successful starting points for solvent flattening.

16.1.9.3. *Integration with multiple-beam diffraction*

Recent experimental work in the field of multiple-beam diffraction provides grounds for hope that a generally applicable solution to the problem of obtaining individual invariant values can be found. It has been shown that triplet invariants can be measured for lysozyme with a mean error of approximately 20° (Weckert *et al.*, 1993; Weckert & Hümmer, 1997). In addition, direct methods strengthened by simulated triplet invariants have been used to redetermine the structure of BPTI at resolutions as low as 2.0 Å (Mathiesen & Mo, 1997, 1998). Currently, the one-at-a-time methods used to measure triplet phases seriously limit practical applications, but faster methods of data collection have been proposed (Shen, 1998). If the means can, in fact, be found for measuring significant numbers of triplet phases quickly and accurately, dual-space direct methods may become routinely applicable to much lower resolution data than is currently possible.

Acknowledgements

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 HCM Institutional Grant ERB CHBG CT 940731 from the European Commission. GMS and IU wish to thank Thammarat Aree, Zbigniew Dauter, Judith Flippen-Anderson, Carlos Frazão, Jörg Kärcher, Katrin Gessler, Håkon Hope, Victor Lamzin, David Langs, Lukatz Lebioda, Paolo Lubini, Peer Mittl, Emilio Parisini, Erich Paulus, Ehmke Pohl, Thierry Prange, Joe Reibenspiess, Martina Schäfer, Thomas Schneider, Markus Teichert, László Vértesy and Martin Walsh for discussions and/or generously providing data for structures referred to in this manuscript. The authors would also like to thank Melda Tugac, Gloria Del Bel and Sandra Finken, who assisted in the preparation of the manuscript.