International Tables for Crystallography (2006). Vol. F, Chapter 16.1, pp. 333–345.

16. DIRECT METHODS

16.1. *Ab initio* phasing

By G. M. Sheldrick, H. A. Hauptman, C. M. Weeks, R. Miller and I. Usón

16.1.1. Introduction

Ab initio methods for solving the crystallographic phase problem rely on diffraction amplitudes alone and do not require prior knowledge of any atomic positions. General features that are not specific to the structure in question (*e.g.* the presence of disulfide bridges or solvent regions) can, however, be utilized. For the last three decades, most small-molecule structures have been routinely solved by *direct methods*, a class of *ab initio* methods in which probabilistic phase relations are used to derive reflection phases from the measured amplitudes. Direct methods, implemented in widely used highly automated computer programs such as *MULTAN* (Main *et al.*, 1980), *SHELXS* (Sheldrick, 1990), *SAYTAN* (Debaerdemaeker *et al.*, 1985) and *SIR* (Burla *et al.*, 1989), provide computationaly efficient solutions for structures containing fewer than approximately 100 independent non-H atoms. However, larger structures are not consistently amenable to these programs and, in fact, few unknown structures with more than 200 independent equal atoms have ever been solved using these programs.

Successful applications to native data for structures that could legitimately be regarded as small macromolecules awaited the development of a direct-methods procedure (Weeks *et al.*, 1993) that has come to be known as *Shake-and-Bake*. The distinctive feature of this procedure is the repeated and unconditional alternation of reciprocal-space phase refinement (*Shaking*) with a complementary real-space process that seeks to improve phases by applying constraints (*Baking*). Consequently, it yields a computer-intensive algorithm, requiring two Fourier transformations during each cycle, which has been made feasible in recent years due to the tremendous increases in computer speed. The first previously unknown structures determined by *Shake-and-Bake* were two forms of the 100-atom peptide ternatin (Miller *et al.*, 1993). Subsequent applications of the *Shake-and-Bake* algorithm have involved structures containing as many as 2000 independent non-H atoms (Frazão *et al.*, 1999) provided that accurate diffraction data have been measured to a resolution of 1.2 Å or better.

The basic theory underlying direct methods has been summarized in an excellent chapter (Giacovazzo, 2001) in *IT* B to which the reader is referred for details. The present chapter focuses on those aspects of direct methods that have proven useful for larger molecules (more than 250 independent non-H atoms) or are unique to the macromolecular field. These include direct-methods applications that utilize anomalous-dispersion measurements or multiple diffraction patterns [*i.e.*, single isomorphous replacement (SIR), single anomalous scattering (SAS) and multiple-wavelength data]. The easiest way to combine isomorphous or anomalous-scattering information with direct methods is to first compute difference structure factors and then to apply direct methods to the difference data. Using this approach, the dual-space *Shake-and-Bake* procedure has been used to solve the anomalously scattering substructure of the selenomethionine derivative of an epimerase enzyme that has 70 selenium sites (Deacon & Ealick, 1999). Substructure applications require only the 2.5–3.0 Å data normally included in multiple wavelength anomalous dispersion (MAD) measurements, and data sets truncated even to 5 Å have led to solutions.

A formal integration of the probabilistic machinery of direct methods with isomorphous replacement and anomalous dispersion was initiated in 1982 (Hauptman, 1982*a*,*b*). Although practical applications of this and subsequent related theory have been limited so far, such applications are likely to have greater importance in the future, and progress is described in Sections 16.1.9.1 and 16.1.9.2. Similarly, the combination of direct methods with multiple-beam diffraction is still in its infancy. However, preliminary studies indicate that the information gleaned from multiple-beam data will greatly strengthen existing techniques (Weckert *et al.*, 1993). Progress in this area is summarized in Section 16.1.9.3.

16.1.2. Normalized structure-factor magnitudes

For purposes of direct-methods computations, the usual structure factors, $F_{\mathbf{H}}$, are replaced by the *normalized structure factors* (Hauptman & Karle, 1953),

$$E_{\mathbf{H}} = |E_{\mathbf{H}}| \exp(i\varphi_{\mathbf{H}}),$$

$$|E_{\mathbf{H}}| = \frac{|F_{\mathbf{H}}|}{\left\langle |F_{\mathbf{H}}|^2 \right\rangle^{1/2}} = \frac{k\left\langle \exp\left[-B_{\rm iso}(\sin\theta)^2/\lambda^2\right]\right\rangle^{-1} |F_{\mathbf{H}}|_{\rm meas}}{\left(\varepsilon_{\mathbf{H}} \sum_{j=1}^{N} f_j^2\right)^{1/2}}, \qquad (16.1.2.1)$$

where the angle brackets indicate probabilistic or statistical expectation values, the $|E_{\mathbf{H}}|$ and $|F_{\mathbf{H}}|$ are structure-factor magnitudes, the $\varphi_{\mathbf{H}}$ are the corresponding phases, k is the absolute scaling factor for the measured magnitudes, $B_{\rm iso}$ is an overall isotropic atomic mean-square displacement parameter, the f_j are the atomic scattering factors for the N atoms in the unit cell, and the $\varepsilon_{\mathbf{H}} \geq 1$ are factors that account for multiple enhancement of the average intensities for certain special reflection classes due to space-group symmetry (Shmueli & Wilson, 2001). The condition $\langle |E|^2 \rangle = 1$ is always imposed. Unlike $\langle |F_{\mathbf{H}}| \rangle$, which decreases as $\sin(\theta)/\lambda$ increases, the values of $\langle |E_{\mathbf{H}}| \rangle$ are constant for concentric resolution shells. Thus, the normalization process places all reflections on a common basis, and this is a great advantage with regard to the probability distributions that form the foundation for direct methods. Normalizing a set of reflections by means of equation (16.1.2.1) does not require any information about atomic positions. However, if some structural information, such as the configuration, orientation, or position of certain atomic groupings, is available, then this information can be applied to obtain a better model for the expected intensity distribution (Main, 1976). The distribution of $|E|$ values is, in principle and often in practice, independent of the unit-cell size and contents, but it does depend on whether a centre of symmetry is present, as shown in Table 16.1.2.1.

Direct-methods applications having the objective of locating SIR or SAS substructures require the computation of normalized *difference* structure-factor magnitudes, $|E_\Delta|$. This can, for example, be accomplished with the following series of programs from Blessing's data-reduction and error-analysis routines (*DREAR*): *LEVY* and *EVAL* for structure-factor normalization as specified by equation (16.1.2.1) (Blessing *et al.*, 1996), *LOCSCL* for local scaling of the SIR and SAS magnitudes (Matthews & Czerwinski, 1975; Blessing, 1997), and *DIFFE* for computing the actual difference magnitudes (Blessing & Smith, 1999). The *SnB* program

Table 16.1.2.1. *Theoretical values pertaining to* $|E|$*'s*

	Centrosymmetric	Noncentrosymmetric
Average $|E|^2$	1.000	1.000
Average $||E^2| - 1|$	0.968	0.736
Average $|E|$	0.798	0.886
$|E| > 1$ (%)	32.0	36.8
$|E| > 2$ (%)	5.0	1.8
$|E| > 3$ (%)	0.3	0.01

(see Section 16.1.7) provides a convenient interface to the *DREAR* suite.

16.1.2.1. *SIR differences*

Given the individual normalized structure-factor magnitudes $(|E_{\rm nat}|, |E_{\rm der}|)$ and the atomic scattering factors $|f_j| = |f_j^0 + f_j' + if_j''| = [(f_j^0 + f_j')^2 + (f_j'')^2]^{1/2}$ which allow for the possibility of anomalous scattering, then greatest-lower-bound estimates of SIR difference-E magnitudes are

$$|E_\Delta| = \frac{\left| \left(\sum_{j=1}^{N_{\rm der}} |f_j|^2\right)^{1/2} |E_{\rm der}| - \left(\sum_{j=1}^{N_{\rm nat}} |f_j|^2\right)^{1/2} |E_{\rm nat}| \right|}{q\left[\left(\sum_{j=1}^{N_{\rm der}} |f_j|^2\right) - \left(\sum_{j=1}^{N_{\rm nat}} |f_j|^2\right)\right]^{1/2}}, \tag{16.1.2.2}$$

where $q = q_0 \exp(q_1 s^2 + q_2 s^4)$ is a least-squares-fitted empirical renormalization scaling function, dependent on $s = \sin(\theta)/\lambda$, that imposes the condition $\langle |E_\Delta|^2 \rangle = 1$ and serves to define q_0, q_1 and q_2.

16.1.2.2. *SAS differences*

Given Friedel pairs of normalized structure-factor magnitudes $(|E_{+\mathbf{H}}|, |E_{-\mathbf{H}}|)$ and the atomic scattering factors, then the greatest-lower-bound estimates of SAS difference $|E|$'s are

$$|E_\Delta| = \frac{\left[\sum_{j=1}^{N} (f_j^0 + f_j')^2 + (f_j'')^2\right]^{1/2} ||E_{+\mathbf{H}}| - |E_{-\mathbf{H}}||}{2q\left[\sum_{j=1}^{N} (f_j'')^2\right]^{1/2}}, \tag{16.1.2.3}$$

where, again, q is an empirical renormalization scaling function that imposes the condition $\langle |E_\Delta|^2 \rangle = 1$.

16.1.3. Starting the phasing process

The phase problem of X-ray crystallography may be defined as the problem of determining the phases φ of the normalized structure factors E when only the magnitudes $|E|$ are given. Owing to the atomicity of crystal structures and the redundancy of the known magnitudes, the phase problem is overdetermined and is, therefore, solvable in principle. This overdetermination implies the existence of relationships among the E's and, since the magnitudes $|E|$ are presumed to be known, the existence of identities among the phases that are dependent on the known magnitudes alone. The techniques of probability theory lead to the joint probability distributions of arbitrary collections of E from which the conditional probability distributions of selected sets of phases, given the values of suitably chosen magnitudes $|E|$, may be inferred.

16.1.3.1. *Structure invariants*

The magnitude-dependent entities that constitute the foundation of direct methods are linear combinations of phases called *structure invariants*. The term 'structure invariant' stems from the fact that the values of these quantities are independent of the choice of origin. The most useful of the structure invariants are the three-phase or *triplet invariants*,

$$\Phi_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \tag{16.1.3.1}$$

the conditional probability distribution (Cochran, 1955), given $A_{\mathbf{HK}}$, of which is

$$P(\Phi_{\mathbf{HK}}) = [2\pi I_0(A_{\mathbf{HK}})]^{-1} \exp(A_{\mathbf{HK}} \cos \Phi_{\mathbf{HK}}), \tag{16.1.3.2}$$

where

$$A_{\mathbf{HK}} = (2/N^{1/2})|E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{H}+\mathbf{K}}| \tag{16.1.3.3}$$

and N is the number of atoms, here presumed to be identical, in the asymmetric unit of the corresponding primitive unit cell. This distribution is illustrated in Fig. 16.1.3.1. The expected value of the cosine of a particular triplet, $\Phi_{\mathbf{HK}}$, is given by the ratio of modified Bessel functions, $I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})$.

Estimates of the invariant values are most reliable when the normalized structure-factor magnitudes ($|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$ and $|E_{-\mathbf{H}-\mathbf{K}}|$) are large and the number of atoms in the unit cell, N, is small. This is the primary reason why direct phasing is more difficult for macromolecules than it is for small molecules. Four-phase or quartet invariants have proven helpful in small-molecule structure determination, particularly when used passively as the basis for a figure of merit (DeTitta *et al.*, 1975). However, the reliability of these invariants, as given by their conditional probability distribution (Hauptman, 1975), is proportional to $1/N$, and they have not as yet been shown to be useful for macromolecular phasing. The reliability of higher-order invariants decreases even more rapidly as structure size increases.

16.1.3.2. *'Multisolution' methods and trial structures*

Successful crystal structure determination requires that sufficient phases be found such that a Fourier map computed using the corresponding structure factors will reveal the atomic positions. It is particularly important that the biggest terms (*i.e.*, largest $|E|$) be included in the Fourier series. Thus, the first step in the phasing

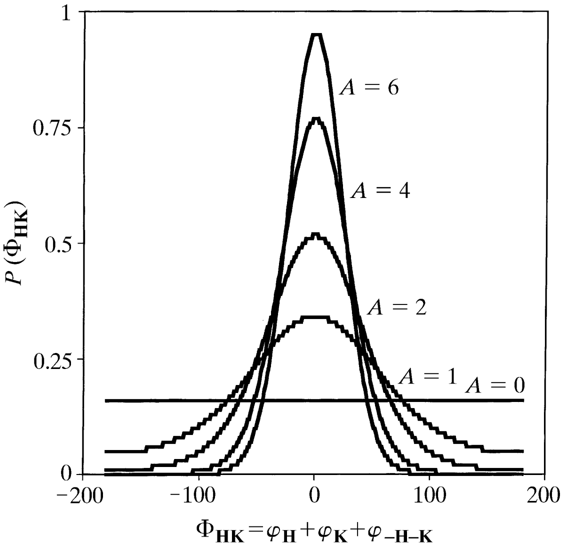


Fig. 16.1.3.1. The conditional probability distribution of the three-phase structure invariants.

process is to sort the reflections in decreasing order according to their $|E|$ values and to choose the number of large $|E|$ reflections that are to be phased. The second step is to generate the possible invariants involving these intense reflections and then to sort them in decreasing order according to their $A_{\mathbf{HK}}$ values. Those invariants with the largest $A_{\mathbf{HK}}$ values are retained in sufficient number to achieve the desired overdetermination. *Ab initio* phase determination by direct methods requires not only a set of invariants, the average values of the cosines of which are presumed to be known, but also a set of starting phases. Therefore, the third step in the phasing process is the assignment of initial phase values. If enough pairs of phases, $\varphi_{\mathbf{K}}$ and $\varphi_{-\mathbf{H}-\mathbf{K}}$, are known, the structure invariants can then be used to generate further phases ($\varphi_{\mathbf{H}}$) which, in turn, can be used to evaluate still more phases. Repeated iterations will permit most reflections with large $|E_{\mathbf{H}}|$ to be phased.

Depending on the space group, a small number of phases can be assigned arbitrarily in order to fix the origin position and, in noncentrosymmetric space groups, the enantiomorph. However, except for the simplest structures, these reflections provide an inadequate foundation for further phase development. Consequently, a 'multisolution' or multi-trial approach (Germain & Woolfson, 1968) is normally taken in which other reflections are each assigned many different starting values in the hope that one or more of the resultant phase combinations will lead to a solution. Solutions, if they occur, must be identified on the basis of some suitable figure of merit. Although phases can be evaluated sequentially, the order determined by a so-called convergence map (Germain *et al.*, 1970), it has become standard in recent years to use a random-number generator to assign initial values to all available phases from the outset (Baggio *et al.*, 1978; Yao, 1981). A variant of this procedure is to use the random-number generator to assign initial coordinates to the atoms in the trial structures and then to obtain initial phases from a structure-factor calculation.

16.1.4. Reciprocal-space phase refinement or expansion (*shaking*)

Once a set of initial phases has been chosen, it must be refined against the set of structure invariants whose values are presumed known. In theory, any of a variety of optimization methods could be used to extract phase information in this way. However, so far only two (tangent refinement and parameter-shift optimization of the minimal function) have been shown to be of practical value.

16.1.4.1. *The tangent formula*

The *tangent formula*,

$$\tan(\varphi_{\mathbf{H}}) = \frac{-\sum_{\mathbf{K}} |E_{\mathbf{K}}E_{-\mathbf{H}-\mathbf{K}}| \sin(\varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} |E_{\mathbf{K}}E_{-\mathbf{H}-\mathbf{K}}| \cos(\varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}})}, \qquad (16.1.4.1)$$

(Karle & Hauptman, 1956), is the relationship used in conventional direct-methods programs to compute $\varphi_{\mathbf{H}}$ given a sufficient number of pairs ($\varphi_{\mathbf{K}}, \varphi_{-\mathbf{H}-\mathbf{K}}$) of known phases. It can also be used within the phase-refinement portion of the dual-space *Shake-and-Bake* procedure (Weeks, Hauptman *et al.*, 1994; Sheldrick & Gould, 1995). The variance associated with $\varphi_{\mathbf{H}}$ depends on $\sum_{\mathbf{K}} E_{\mathbf{H}}E_{\mathbf{K}}E_{-\mathbf{H}-\mathbf{K}}/N^{1/2}$ and, in practice, the estimate is only reliable for $|E_{\mathbf{H}}| \gg 1$ and for structures with a limited number of atoms (N). If equation (16.1.4.1) is used to redetermine previously known phases, the phasing process is referred to as *tangent-formula refinement*; if only new phases are determined, the phasing process is *tangent expansion*.

The tangent formula can be derived using the assumption of equal resolved atoms. Nevertheless, it suffers from the disadvantage that, in space groups without translational symmetry, it is perfectly fulfilled by a false solution with all phases equal to zero, thereby giving rise to the so-called 'uranium-atom' solution with one dominant peak in the corresponding Fourier synthesis. In conventional direct-methods programs, the tangent formula is often modified in various ways to include (explicitly or implicitly) information from the so-called 'negative' quartet invariants (Schenk, 1974; Hauptman, 1974; Giacovazzo, 1976) that are dependent on the smallest as well as the largest E magnitudes. Such modified tangent formulas do indeed largely overcome the problem of pseudosymmetric solutions for small N, but because of the dependence of quartet-term probabilities on $1/N$, they are little more effective than the normal tangent formula for large N.

16.1.4.2. *The minimal function*

Constrained minimization of an objective function like the *minimal function*,

$$R(\Phi) = \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \{\cos \Phi_{\mathbf{HK}} - [I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})]\}^2 \Big/ \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \qquad (16.1.4.2)$$

(Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.*, 1994), provides an alternative approach to phase refinement or phase expansion. $R(\Phi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and the expected values of the same triplets as given by the ratio of modified Bessel functions. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph (the minimal principle). Experimentation has thus far confirmed that, when the minimal function is used actively in the phasing process and solutions are produced, the final trial structure corresponding to the smallest value of $R(\Phi)$ is a solution provided that $R(\Phi)$ is calculated directly from the atomic positions before the phase-refinement step (Weeks, DeTitta *et al.*, 1994). Therefore, $R(\Phi)$ is also an extremely useful figure of merit. The minimal function can also include contributions from higher-order (*e.g.* quartet) invariants, although their use is not as imperative as with the tangent formula because the minimal function does not have a minimum when all phases are zero. In practice, quartets are rarely used in the minimal function because they increase the CPU time while adding little useful information for large structures. The cosine function in equation (16.1.4.2) can also be replaced by other functions of the phases giving rise to alternative minimal functions. In particular, an exponential expression has been found to give superior results for several $P1$ structures (Hauptman *et al.*, 1999).

16.1.4.3. *Parameter shift*

In principle, any minimization technique could be used to minimize $R(\Phi)$ by varying the phases. So far, a seemingly simple algorithm, known as parameter shift (Bhuiya & Stanley, 1963), has proven to be quite powerful and efficient as an optimization method when used within the *Shake-and-Bake* context to reduce the value of the minimal function. For example, a typical phase-refinement stage consists of three iterations or scans through the reflection list, with each phase being shifted a maximum of two times by 90° in either the positive or negative direction during each iteration. The refined value for each phase is selected, in turn, through a process which involves evaluating the minimal function using the original phase and each of its shifted values (Weeks, DeTitta *et al.*, 1994). The phase value that results in the lowest minimal-function value is chosen at each step. Refined phases are used immediately in the subsequent refinement of other phases. It should be noted that the parameter-shift routine is similar to that used in ψ-map refinement

(White & Woolfson, 1975) and *XMY* (Debaerdemaeker & Woolfson, 1989).

16.1.5. Real-space constraints (*baking*)

Peak picking is a simple but powerful way of imposing an atomicity constraint. The potential for real-space phase improvement in the context of small-molecule direct methods was recognized by Jerome Karle (1968). He found that even a relatively small, chemically sensible, fragment extracted by manual interpretation of an electron-density map could be expanded into a complete solution by transformation back to reciprocal space and then performing additional iterations of phase refinement with the tangent formula. Automatic real-space electron-density map interpretation in the *Shake-and-Bake* procedure consists of selecting an appropriate number of the largest peaks in each cycle to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance between atoms. If markedly unequal atoms are present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space in a subsequent structure-factor calculation. Thus, *a priori* knowledge concerning the chemical composition of the crystal is utilized, but no knowledge of constitution is required or used during peak selection. It is useful to think of peak picking in this context as simply an extreme form of density modification appropriate when atomic resolution data are available. In theory, under appropriate conditions it should be possible to substitute alternative density-modification procedures such as low-density elimination (Shiono & Woolfson, 1992; Refaat & Woolfson, 1993) or solvent flattening (Wang, 1985), but no practical applications of such procedures have yet been made. The imposition of physical constraints counteracts the tendency of phase refinement to propagate errors or produce overly consistent phase sets. Several variants of peak picking, which are discussed below, have been successfully employed within the framework of *Shake-and-Bake*.

16.1.5.1. *Simple peak picking*

In its simplest form, peak picking consists of simply selecting the top N_u E-map peaks where N_u is the number of unique non-H atoms in the asymmetric unit. This is adequate for true small-molecule structures. It has also been shown to work well for heavy-atom or anomalously scattering substructures where N_u is taken to be the number of expected substructure atoms (Smith *et al.*, 1998; Turner *et al.*, 1998). For larger structures ($N_u > 100$), it is likely to be better to select about $0.8N_u$ peaks, thereby taking into account the probable presence of some atoms that, owing to high thermal motion or disorder, will not be visible during the early stages of a structure determination. Furthermore, a recent study (Weeks & Miller, 1999*b*) has shown that structures in the 250–1000-atom range which contain a half dozen or more moderately heavy atoms (*i.e.*, S, Cl, Fe) are more easily solved if only $0.4N_u$ peaks are selected. The only chemical information used at this stage is a minimum inter-peak distance, generally taken to be 1.0 Å. For substructure applications, a larger minimum distance (*e.g.* 3 Å) is more appropriate.

16.1.5.2. *Iterative peaklist optimization*

An alternative approach to peak picking is to select approximately N_u peaks as potential atoms and then eliminate some of them, one by one, while maximizing a suitable figure of merit such as

$$P = \sum_{\mathbf{H}} |E_c^2|(|E_o^2| - 1). \qquad (16.1.5.1)$$

The top N_u peaks are used as potential atoms to compute $|E_c|$. The atom that leaves the highest value of P is then eliminated. Typically, this procedure, which has been termed *iterative peaklist optimization* (Sheldrick & Gould, 1995), is repeated until only $2N_u/3$ atoms remain. Use of equation (16.1.5.1) may be regarded as a reciprocal-space method of maximizing the fit to the origin-removed sharpened Patterson function, and it is used for this purpose in molecular replacement (Beurskens, 1981). Subject to various approximations, maximum-likelihood considerations also indicate that it is an appropriate function to maximize (Bricogne, 1998). Iterative peaklist optimization provides a higher percentage of solutions than simple peak picking, but it suffers from the disadvantage of requiring much more CPU time.

16.1.5.3. *Random omit maps*

A third peak-picking strategy also involves selecting approximately N_u of the top peaks and eliminating some, but, in this case, the deleted peaks are chosen at random. Typically, one-third of the potential atoms are removed, and the remaining atoms are used to compute E_c. By analogy to the common practice in macromolecular crystallography of omitting part of a structure from a Fourier calculation in hopes of finding an improved position for the deleted fragment, this version of peak picking is described as making a *random omit map*. This procedure is a little faster than simply picking N_u atoms because fewer atoms are used in the structure-factor calculation. More important is the fact that, like iterative peaklist optimization, it has the potential for being a more efficient search algorithm.

16.1.6. Fourier refinement (*twice baking*)

E-map recycling, but without phase refinement (Sheldrick, 1982, 1990; Kinneging & de Graaff, 1984), has been frequently used in conventional direct-methods programs to improve the completeness of the solutions after phase refinement. It is important to apply Fourier refinement to *Shake-and-Bake* solutions also because such processing significantly increases the number of resolved atoms, thereby making the job of map interpretation much easier. Since phase refinement *via* either the tangent formula or the minimal function requires relatively accurate invariants that can only be generated using the larger E magnitudes, a limited number of reflections are phased during the actual dual-space cycles. Working with a limited amount of data has the added advantage that less CPU time is required. However, if the current trial structure is the 'best' so far based on a figure of merit (either the minimal function or a real-space criterion), then it makes sense to subject this structure to Fourier refinement using additional data, thereby reducing series-termination errors. The correlation coefficient

$$\begin{aligned}\mathrm{CC} = {}&\left[\left(\sum wE_o^2E_c^2 \sum w\right) - \left(\sum wE_o^2 \sum wE_c^2\right)\right]\\ &\times \left\{\left[\left(\sum wE_o^4 \sum w\right) - \left(\sum wE_o^2\right)^2\right]\right.\\ &\times \left.\left[\left(\sum wE_c^4 \sum w\right) - \left(\sum wE_c^2\right)^2\right]\right\}^{-1/2} \qquad (16.1.6.1)\end{aligned}$$

(Fujinaga & Read, 1987), where weights $w = 1/[0.04 + \sigma^2(E_o)]$, has been found to be an especially effective figure of merit when used with all the data and is, therefore, suited for identifying the most promising trial structure at the end of Fourier refinement. Either simple peak picking or iterative peaklist optimization can be employed during the Fourier-refinement cycles in conjunction with weighted E maps (Sim, 1959). The final model can be further improved by isotropic displacement parameter ($B_{\rm iso}$) refinement for the individual atoms (Usón *et al.*, 1999) followed by calculation of the Sim (1959) or sigma-A (Read, 1986) weighted map. This is particularly useful when the requirement of atomic resolution is barely fulfilled, and it makes it easier to interpret the resulting maps by classical macromolecular methods.

16.1.7. Computer programs for dual-space phasing

The *Shake-and-Bake* algorithm has been implemented independently in two computer programs. These are (1) *SnB* written in Buffalo at the Hauptman–Woodward Institute, principally by Charles Weeks and Russ Miller (Miller *et al.*, 1994; Weeks & Miller, 1999*a*), and (2) *SHELXD* (which is also known by the alias '*Halfbaked*'), written in Göttingen by George Sheldrick (Sheldrick, 1997, 1998). *SHELXD* attempts to do more during the real-space (*baking*) stage than is available to the user with the current version of *SnB*. The most recent public release of *SnB* is available at http://www.hwi.buffalo.edu/SnB/ along with documentation, test data and other pertinent information. *SHELXD* will be released when testing is complete; for details see the *SHELX* homepage at http://shelx.uni-ac.gwdg.de/SHELX/.

16.1.7.1. *Flowchart and program comparison*

A flowchart for the generic *Shake-and-Bake* algorithm, which provides the foundation for both programs, is presented in Fig. 16.1.7.1. It contains two refinement loops embedded in the trial-structure loop. The first of these loops (steps 5–9) is a dual-space phase-improvement loop entered by all trial structures, and the second (steps 11–14) is a real-space Fourier-refinement loop entered only by those trial structures that are currently judged to be the best on the basis of some figure of merit. These loops have been called the internal and external loops, respectively, in previous descriptions of the *SHELXD* program (*e.g.* Sheldrick & Gould, 1995; Sheldrick, 1997, 1998). Currently, the major algorithmic differences between the programs are the following:

(*a*) During the reciprocal-space segment of the dual-space loop (Fig. 16.1.7.1, step 5), *SnB* can perform tangent refinement or use parameter shift to reduce the minimal function [equation (16.1.4.2)] or an exponential variant of the minimal function (Hauptman *et al.*, 1999). *SHELXD* can perform either Karle-type tangent expansion (Karle, 1968) or parameter-shift refinement based on either the minimal function or the tangent formula. During tangent or parameter-shift refinement, all phases computed in the preceding structure-factor calculation (step 4 or 9) are refined. During tangent expansion in *SHELXD*, the phases of (typically) the 40% highest calculated *E* magnitudes are held fixed, and the phases of the remaining 60% are determined by using the tangent formula.

(*b*) In real space, *SnB* uses simple peak picking, varying the number of peaks selected on the basis of structure size and composition. *SHELXD* contains provisions for all the forms of peak picking described above.

(*c*) *SnB* relies primarily on the minimal function [equation (16.1.4.2)] as a figure of merit whereas *SHELXD* uses the correlation coefficient [equation (16.1.6.1)], calculated using all data, after the final dual-space (internal) cycle and in the real-space (external) loop.

16.1.7.2. *Parameters and procedures*

All of the major parameters of the *Shake-and-Bake* procedure (*i.e.*, the numbers of refinement cycles, phases, triplet invariant relationships and peaks selected) are a function of structure size and can be expressed in terms of N_u, the number of unique non-H atoms in the asymmetric unit. These parameters have been fine-tuned in a series of tests using data for both small and large molecules (Weeks, DeTitta *et al.*, 1994; Chang *et al.*, 1997; Weeks & Miller, 1999*b*). Default (recommended) parameter values used in the *SnB* program are summarized in Table 16.1.7.1. At resolutions in the 1.1–1.4 Å range, recalcitrant data sets can sometimes be made to yield solutions if (1) the phase:invariant ratio is increased from 1:10 to values ranging between 1:20 and 1:50 or (2) the number of dual-space refinement cycles is doubled or tripled. The presence of moderately heavy atoms (*e.g.* S, C, Fe) greatly increases the probability of success at resolutions less than 1.2 Å; in general, the higher the fraction of such atoms the more the resolution requirement can be relaxed, provided that these atoms have low *B* values. Thus, disulfide bridges are much more helpful than methionine sulfur atoms because they tend to have lower *B* values. Parameter recommendations for substructures are based on an analysis of the peak-wavelength anomalous-difference data for S-adenosylhomocysteine (AdoHcy) hydrolase (Turner *et al.*, 1998). Parameter shift with a maximum of two 90° steps [indicated by the shorthand notation PS(90°, 2)] is the default phase-refinement mode. However, some structures (especially large *P*1 structures) may respond better to a single larger shift [*e.g.* PS(157.5°, 1)]

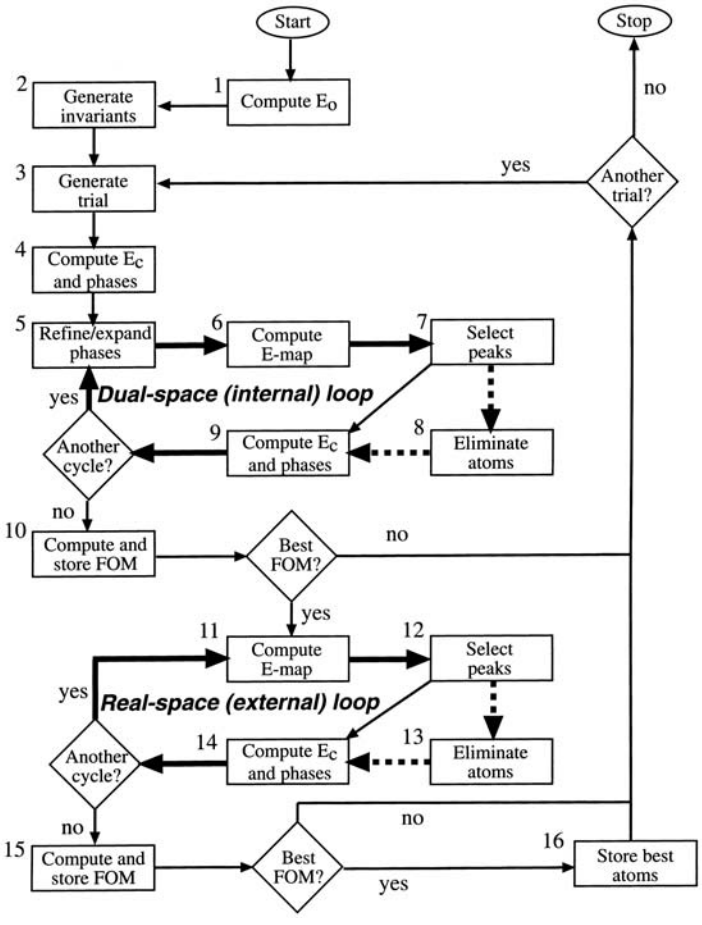


Fig. 16.1.7.1. A flowchart for the *Shake-and-Bake* procedure, which is implemented in both *SnB* and *SHELXD*. The essence of the method is the dual-space approach of refining trial structures as they shuttle between real and reciprocal space. In the general case, steps 7 and 12 are any density-modification procedure, and steps 9 and 14 are inverse Fourier transforms rather than structure-factor calculations. The optional steps 8 and 13 take the form of *iterative peaklist optimization* or *random omit maps* in *SHELXD*. Any suitable starting model can be used in step 3, and *SHELXD* attempts to improve on random models (when possible) by utilizing Patterson-based information. Step 4 is bypassed if phase sets (random or otherwise) provide the starting point for the dual-space loop. *SHELXD* enters the real-space loop if the FOM (correlation coefficient) is within a specified threshold (1–5%) of the best value so far.

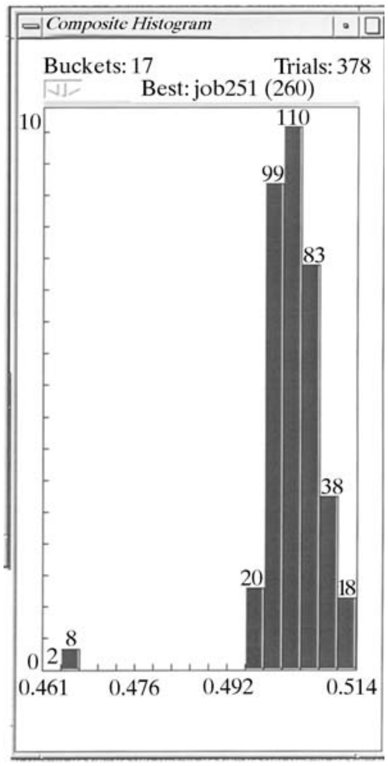
Table 16.1.7.1. *Recommended parameter values for the SnB program*

Values are expressed in terms of N_u, the number of unique non-H atoms (solvent atoms are typically ignored). Full-structure recommendations are for data sets measured to 1.1 Å resolution or better. Only heavy atoms or anomalous scatterers are counted for substructures.

Parameter	Full structures	Substructures
Phases	$10N_u$	$30N_u$
Triplet invariants	$100N_u$	$300N_u$
Peaks (with S, Cl) Peaks (no 'heavy')	$0.4N_u$ $0.8N_u$	N_u
Cycles	$N_u/2$ if $N_u < 100$ or if $N_u < 400$ with S, Cl *etc.*; N_u otherwise	$2N_u$ (minimum 20)

(Deacon *et al.*, 1998). This seems to reduce the frequency of false minima (see Section 16.1.8.2).

In general, the parameter values used in *SHELXD* are similar to those used in *SnB*. However, the *combination* of random omit maps with tangent extension has been found to be the most effective strategy within the context of *SHELXD*. Consequently, it is used as the default operational mode (see Section 16.1.8.4 for details).



Composite Histogram

Buckets: 17 Trials: 378

Best: job251 (260)

Fig. 16.1.7.2. A histogram of figure-of-merit values (minimal function) for 378 scorpion toxin II trials. This bimodal histogram suggests that ten trials are solutions.

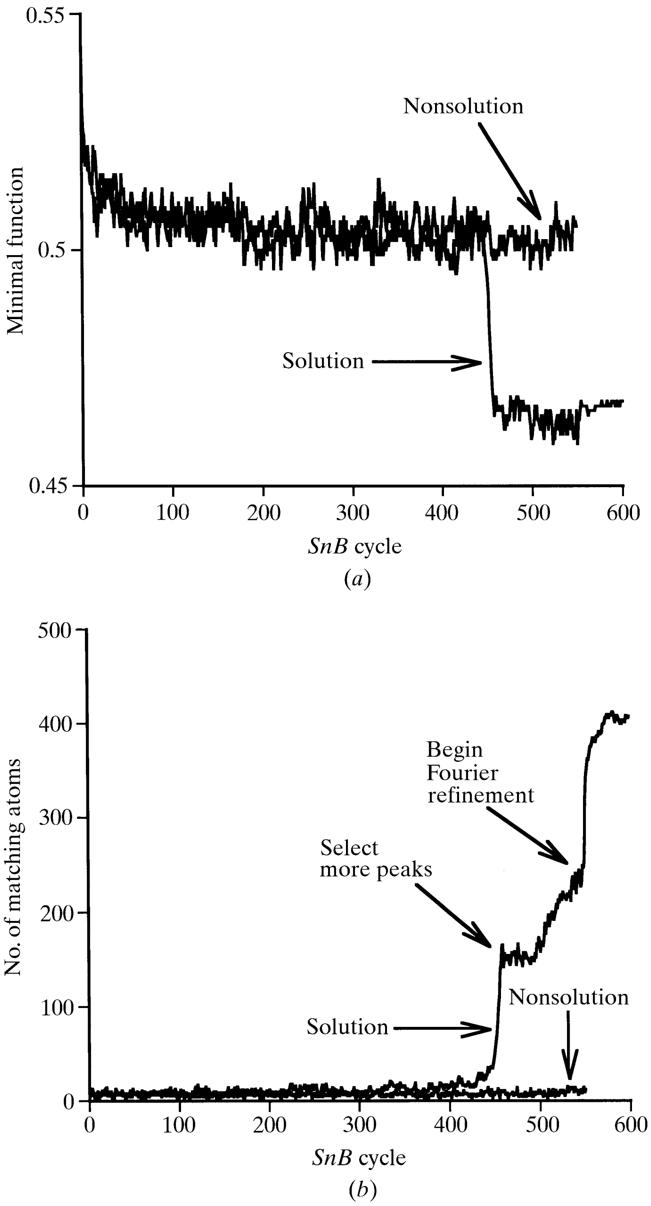


Fig. 16.1.7.3. Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function *of Shake-and-Bake* cycle. (*a*) Minimal-function figure of merit, and (*b*) number of peaks closer than 0.5 Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 500 (N_u) cycles, and 500 peaks (N_u) were then selected for an additional 50 ($0.1N_u$) dual-space cycles. The solution (which had the lowest minimal-function value) was then subjected to 50 cycles of Fourier refinement.

16.1.7.3. *Recognizing solutions*

On account of the intensive nature of the computations involved, *SnB* and *SHELXD* are designed to run unattended for long periods while also providing ways for the user to check the status of jobs in progress. The progress of current *SnB* jobs can be followed by monitoring a figure-of-merit histogram for the trial structures that have been processed (Fig. 16.1.7.2). A clear bimodal distribution of figure-of-merit values is a strong indication that a solution has, in fact, been found. However, not all solutions are so obvious, and it sometimes pays to inspect the best trial even when the histogram is unimodal. The course of a typical solution as a function of *SnB* cycle is contrasted with that of a nonsolution in Fig. 16.1.7.3. Minimal-function values for a solution usually decrease abruptly over the course of just a few cycles, and a tool is provided within *SnB* that allows the user to visually inspect the trace of minimal-function values for the best trial completed so far. Fig. 16.1.7.3 shows that the abrupt decrease in minimal-function values corresponds to a simultaneous abrupt increase in the number of peaks close to true atomic positions. In this example, a second abrupt increase in correct peaks occurs when Fourier refinement is started.

Since the correlation coefficient is a relatively absolute figure of merit (given atomic resolution, values greater than 65% almost invariably correspond to correct solutions), it is usually clear when *SHELXD* has solved a structure. The current version of *SHELXD* includes an option for calculating it using the full data every 10 or 20 internal loop cycles, and jumping to the external loop if the value is high enough. Recalculating it every cycle would be computationally less efficient overall.

16.1.8. Applying dual-space programs successfully

The solution of the (known) structure of triclinic lysozyme by *SHELXD* and shortly afterwards by *SnB* (Deacon *et al.*, 1998) finally broke the 1000-atom barrier for direct methods (there happen to be 1001 protein atoms in this structure!). Both programs have also solved a large number of previously unsolved structures that had defeated conventional direct methods; some examples are listed in Table 16.1.8.1. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier-

Table 16.1.8.1. *Some large structures solved by the Shake-and-Bake method*

Previously known test data sets are indicated by an asterisk (*). When two numbers are given in the resolution column, the second indicates the lowest resolution at which truncated data have yielded a solution. The program codes are *SnB* (S) and *SHELXD* (D).

(*a*) Full structures (>300 atoms).

Compound	Space group	N_u (molecule)	N_u + solvent	N_u (heavy)	Resolution (Å)	Program	Reference
Vancomycin	$P4_32_12$	202	258	8Cl	0.9–1.4	S	[1]
			312	6Cl	1.09	D	[2]
Actinomycin X2	$P1$	273	305	—	0.90	D	[3]
Actinomycin Z3	$P2_12_12_1$	186	307	2Cl	0.96	D	[4]
Actinomycin D	$P1$	270	314	—	0.94	D	[4]
Gramicidin A*	$P2_12_12_1$	272	317	—	0.86–1.1	S, D	[5]
DMSO d6 peptide	$P1$	320	326	—	1.20	S	[6]
Er-1 pheromone	$C2$	303	328	7S	1.00	S	[7]
Ristocetin A	$P2_1$	294	420	—	1.03	D	[8]
Crambin*	$P2_1$	327	423	6S	0.83–1.2	S, D	[9], [10]
Hirustasin	$P4_32_12$	402	467	10S	1.2–1.55	D	[11]
Cyclodextrin derivative	$P2_1$	448	467	—	0.88	D	[12]
Alpha-1 peptide	$P1$	408	471	Cl	0.92	S	[13]
Rubredoxin*	$P2_1$	395	497	Fe, 6S	1.0–1.1	S, D	[14]
Vancomycin	$P1$	404	547	12Cl	0.97	S	[15]
BPTI*	$P2_12_12_1$	453	561	7S	1.08	D	[16]
Cyclodextrin derivative	$P2_1$	504	562	28S	1.00	D	[17]
Balhimycin*	$P2_1$	408	598	8Cl	0.96	D	[18]
Mg-complex*	$P1$	576	608	8Mg	0.87	D	[19]
Scorpion toxin II*	$P2_12_12_1$	508	624	8S	0.96–1.2	S	[20]
Amylose-CA26	$P1$	624	771	—	1.10	D	[21]
Mersacidin	$P3_2$	750	826	24S	1.04	D	[22]
Cv HiPIP H42Q*	$P2_12_12_1$	631	837	4Fe	0.93	D	[23]
HEW lysozyme*	$P1$	1001	1295	10S	0.85	S, D	[24], [25]
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20	D	[23]
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20	D	[26]

(*b*) Se substructures (> 25 Se) solved using peak-wavelength anomalous-difference data.

Protein	Space group	Molecular weight (kDa)	Se located	Se total	Resolution (Å)	Program	Reference
SAM decarboxylase	$P2_1$	77	20	26	2.25	S	[27]
AIR synthetase	$P2_12_12_1$	147	28	28	3.0	S	[28]
FTHFS	$R32$	200	28	28	2.5	D	[29]
AdoHcy hydrolase	$C222$	95	30	30	2.8–5.0	S	[30]
Epimerase	$P2_1$	370	64	70	3.0	S	[31]

References: [1] Loll *et al.* (1997); [2] Schäfer *et al.* (1996); [3] Schäfer (1998); [4] Schäfer, Sheldrick, Bahner & Lackner (1998); [5] Langs (1988); [6] Drouin (1998); [7] Anderson *et al.* (1996); [8] Schäfer & Prange (1998); [9] Stec *et al.* (1995); [10] Weeks *et al.* (1995); [11] Usón *et al.* (1999); [12] Aree *et al.* (1999); [13] Prive *et al.* (1999); [14] Dauter *et al.* (1992); [15] Loll *et al.* (1998); [16] Schneider (1998); [17] Reibenspiess (1998); [18] Schäfer, Sheldrick, Schneider & Vértesy (1998); [19] Teichert (1998); [20] Smith *et al.* (1997); [21] Gessler *et al.* (1999); [22] Schneider *et al.* (2000); [23] Parisini *et al.* (1999); [24] Deacon *et al.* (1998); [25] Walsh *et al.* (1998); [26] Frazão *et al.* (1999); [27] Ekstrom *et al.* (1999); [28] Li *et al.* (1999); [29] Radfar *et al.* (2000); [30] Turner *et al.* (1998); [31] Deacon & Ealick (1999).

Table 16.1.8.2. *Overall success rates for full structure solution for hirustasin using different two-atom search vectors chosen from the Patterson peak list*

Resolution (Å)	Two-atom search fragments	Solutions per 1000 attempts
1.2	Top 100 general Patterson peaks	86
1.2	Top 300 general Patterson peaks	38
1.2	One vector, error = 0.08 Å	14
1.2	One vector, error = 0.38 Å	41
1.2	One vector, error = 0.40 Å	219
1.2	One vector, error = 1.69 Å	51
1.4	Top 100 general Patterson peaks	10
1.5	Top 100 general Patterson peaks	4
1.5	One vector, error = 0.29 Å	61

refinement stage. Most of the time, the *Shake-and-Bake* method works remarkably well, even for rather large structures. However, in problematic situations, the user needs to be aware of options that can increase the chance of success.

16.1.8.1. *Utilizing Pattersons for better starts*

When slightly heavier atoms such as sulfur are present, it is possible to start the *Shake-and-Bake* recycling procedure from a set of atomic positions that are consistent with the Patterson function. For large structures, the vectors between such atoms will correspond to Patterson densities around or even below the noise level, so classical methods of locating the positions of these atoms unambiguously from the Patterson are unlikely to succeed. Nevertheless, the Patterson function can still be used to filter sets of starting atoms. This filter is currently implemented as follows in *SHELXD*. First, a sharpened Patterson function (Sheldrick *et al.*, 1993) is calculated, and the top 200 (for example) non-Harker peaks further than a given minimum distance from the origin are selected, in turn, as two-atom translation-search fragments, one such fragment being employed per solution attempt. For each of a large number of random translations, all unique Patterson vectors involving the two atoms and their symmetry equivalents are found and sorted in order of increasing Patterson density. The sum of the smallest third of these values is used as a figure of merit (PMF). Tests showed that although the globally highest PMF for a given two-atom search fragment may not correspond to correct atomic positions, nevertheless, by limiting the number of trials, some correct solutions may still be found. After all the vectors have been used as search fragments (*e.g.* after 200 attempts), the procedure is repeated starting again with the first vector. The two atoms may be used to generate further atoms using a full Patterson superposition minimum function or a weighted difference synthesis (in the current version of *SHELXD*, a combination of the two is used).

In the case of the small protein BPTI (Schneider, 1998), 15 300 attempts based on 100 different search vectors led to four final solutions with mean phase error less than 18°, although none of the globally highest PMF values for any of the search vectors corresponded to correct solutions. Table 16.1.8.2 shows the effect of using different two-atom search fragments for hirustasin, a previously unsolved 55-amino-acid protein containing five disulfide bridges first solved using *SHELXD* (Usón *et al.*, 1999). It is not clear why some search fragments perform so much better than others; surprisingly, one of the more effective search vectors deviates considerably (1.69 Å) from the nearest true S–S vector.

16.1.8.2. *Avoiding false minima*

The frequent imposition of real-space constraints appears to keep dual-space methods from producing most of the false minima that plague practitioners of conventional direct methods. Translated molecules have not been observed (so far), and traditionally problematic structures with polycyclic ring systems and long aliphatic chains are readily solved (McCourt *et al.*, 1996, 1997). False minima of the type that occur primarily in space groups lacking translational symmetry and are characterized by a single large 'uranium' peak do occur frequently in $P1$ and occasionally in other space groups. Triclinic hen egg-white lysozyme exhibits this phenomenon regardless of whether parameter-shift or tangent-formula phase refinement is employed. An example from another space group ($C222$) is provided by the Se substructure data for AdoHcy hydrolase. In this case, many trials converge to false minima if the feature in the *SnB* program that eliminates peaks at special positions is not utilized.

The problem with false minima is most serious if they have a 'better' value of the figure of merit being used for diagnostic purposes than do the true solutions. Fortunately, this is not the case with the uranium 'solutions', which can be distinguished on the basis of the minimal function [equation (16.1.4.2)] or the correlation coefficient [equation (16.1.6.1)]. However, it would be inefficient to compute the latter in each dual-space cycle since it requires that essentially all reflections be used. To be an effective discriminator, the figure of merit must be computed using the phases calculated from the point-atom model, not from the phases directly after refinement. Phase refinement can and does produce sets of phases, such as the uranium phases, which do not correspond to physical reality. Hence, it should not be surprising that such phase sets might appear 'better' than the true phases and could lead to an erroneous choice for the best trial. Peak picking, followed by a structure-factor calculation in which the peaks are sensibly weighted, converts the phase set back to physically allowed values. If the value of the minimal function computed from the refined or *unconstrained* phases is denoted by $R_{\rm unc}$ and the value of the minimal function computed using the *constrained* phases resulting from the atomic model is denoted by $R_{\rm con}$, then a function defined by

$$R \text{ ratio} = (R_{\rm con} - R_{\rm unc})/(R_{\rm con} + R_{\rm unc}) \qquad (16.1.8.1)$$

can be used to distinguish false minima from other nonsolutions as well as the true solutions. Once a trial falls into a false minimum, it never escapes. Therefore, the R ratio can be used, within *SnB*, as a criterion for early termination of unproductive trials. Based on data for several $P1$ structures, it appears that termination of trials with R ratio values exceeding 0.2 will eliminate most false minima without risking rejection of any potential solutions. In the case of triclinic lysozyme, false minima can be recognized, on average, by cycle 25. Since the default recommendation would be for 1000 cycles, a substantial saving in CPU time is realized by using the R ratio early-termination test. It should be noted that *SHELXD* optionally allows early termination of trials if the second peak is less than a specified fraction (*e.g.* 40%) of the height of the first. Generally, but not always, the R-ratio and peak-ratio tests eliminate the same trials.

Recognizing false minima is, of course, only part of the battle. It is also necessary to find a real solution, and essentially 100% of the triclinic lysozyme trials were found to be false minima when the standard parameter-shift conditions of two 90° shifts were used. In fact, significant numbers of solutions occur only when single-shift angles in the range 140–170° are used (Fig. 16.1.8.1), and there is a surprisingly high *success rate* (percentage of trial structures that go to solutions) over a narrow range of angles centred about 157.5°. It is also not surprising that there is a correlated decrease in the percentage of false minima in the range 140–150°. This suggests that a fruitful strategy for structures that exhibit a large percentage

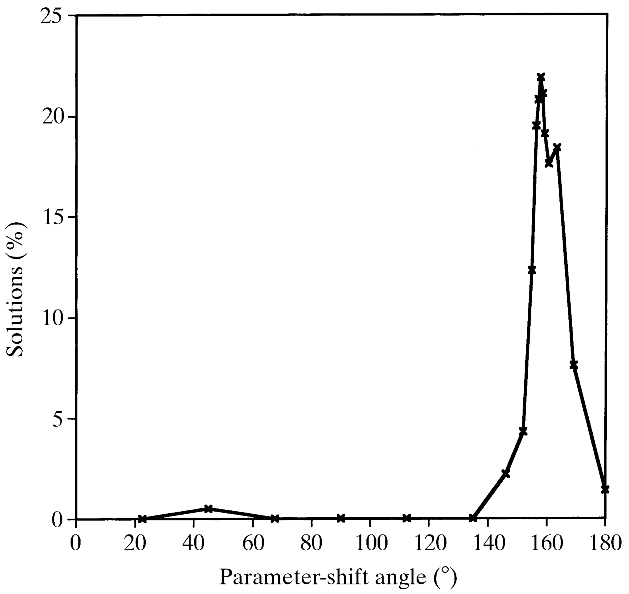


Fig. 16.1.8.1. Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

of false minima would be the following. Run 100 or so trials at each of several shift angles in the range 90–180°, find the smallest angle which gives nearly zero false minima, and then use this angle as a single shift for many trials. Balhimycin is an example of a large non-$P1$ structure that also requires a parameter shift of around 154° to obtain a solution using the minimal function.

16.1.8.3. *Data resolution and completeness*

The importance of the presence of several atoms heavier than oxygen for increasing the chance of obtaining a solution by *SnB* at resolutions less than 1.2 Å was noticed for truncated data from vancomycin and the 289-atom structure of conotoxin EpI (Weeks & Miller, 1999*b*). The results of *SHELXD* application to hirustasin are consistent with this (Usón *et al.*, 1999). The 55-amino-acid protein hirustasin could be solved by *SHELXD* using either 1.2 Å low-temperature data or 1.4 Å room-temperature data; however, as shown in Fig. 16.1.8.2(*a*), the mean phase error (MPE) is significantly better for the 1.2 Å data over the whole resolution range. The MPE is determined primarily by the data-to-parameter ratio, which is reflected in the smaller number of reliable triplet invariants at lower resolution. Although small-molecule interpretation based on peak positions worked well for the 1.2 Å solution (overall MPE = 18°), standard protein chain tracing was required for the 1.4 Å solution (overall MPE = 26°). As is clear from the corresponding electron-density map (Fig. 16.1.8.2*b*), the *Shake-and-Bake* procedure produces easily interpreted protein density even when bonded atoms are barely resolved from each other. The hirustasin structure was also determined with *SHELXD* using 1.55 Å truncated data, and this endeavour currently holds the record for the lowest-resolution successful application of *Shake-and-Bake*.

The relative effects of accuracy, completeness and resolution on *Shake-and-Bake* success rates using *SnB* for three large $P1$ structures were studied by computing error-free data using the known atomic coordinates. The results of these studies, presented in Table 16.1.8.3, show that experimental error contributed nothing of consequence to the low success rates for vancomycin and lysozyme. However, completing the vancomycin data up to the maximum measured resolution of 0.97 Å resulted in a substantial increase in success rate which was further improved to an astounding success rate of 80% when the data were expanded to 0.85 Å.

On account of overload problems, the experimental vancomycin data did not include any data at 10 Å resolution or lower. A total of 4000 reflections were phased in the dual-space loop in the process of solving this structure with the experimental data. Some of these data were then replaced with the largest error-free magnitudes chosen from the missing reflections at several different resolution limits. The results in Table 16.1.8.4 show a tenfold increase in success rate when only 200 of the largest missing magnitudes were supplied, and it made no difference whether these reflections had a maximum resolution of 2.8 Å or were chosen randomly from the whole 0.97 Å sphere. The moral of this story is that, *when collecting data for Shake-and-Bake, it pays to take a second pass using a shorter exposure to fill-in the low-resolution data.*

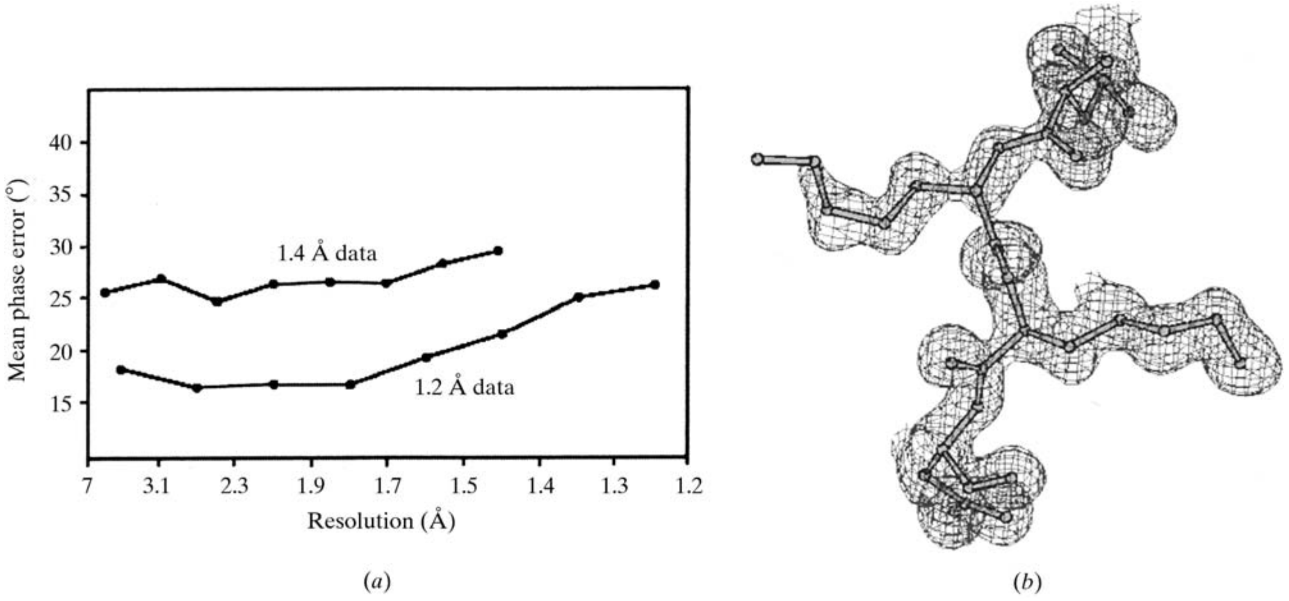


Fig. 16.1.8.2. (*a*) Mean phase error as a function of resolution for the two independent *ab initio SHELXD* solutions of the previously unsolved protein hirustasin. Either the 1.2 Å or the 1.4 Å native data set led to solution of the structure. (*b*) Part of the hirustasin molecule from the 1.4 Å room-temperature data after one round of B-value refinement with fixed coordinates.

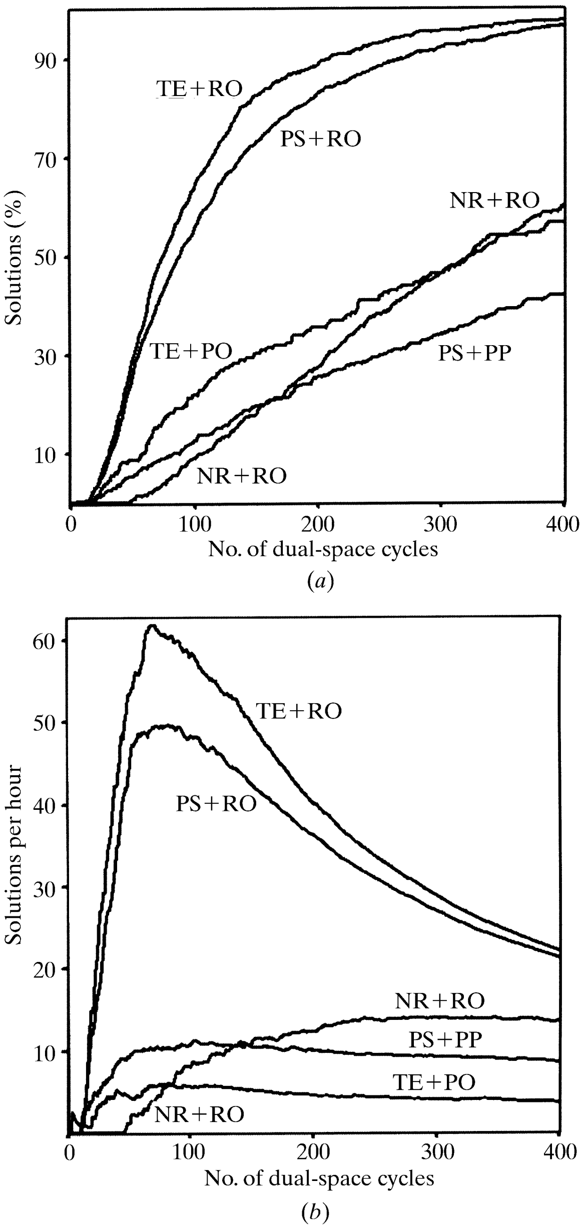
Table 16.1.8.3. *Success rates for three P1 structures illustrate the importance of using complete data to the highest possible resolution*

	Vancomycin	Alpha-1	Lysozyme
Atoms	547	471	~1200
Completeness (%)	80.2	85.6	68.3
Resolution (Å)	0.97	0.90	0.85
Parameter shift	112.5°, 1	90°, 2	90°, 2
Success rates (%)			
Experimental	0.25	14	0
Error-free	0.2	19	0
Error-free complete	14	29	0.8
Error-free complete extended to 0.85 Å	80	42	—

16.1.8.4. *Choosing a refinement strategy*

Variations in the computational details of the dual-space loop can make major differences in the efficacy of *SnB* and *SHELXD*. Recently, several strategies were combined in *SHELXD* and applied to a 148-atom $P1$ test structure (Karle *et al.*, 1989) with the results shown in Fig. 16.1.8.3. The CPU time requirements of parameter-shift (PS) and tangent-formula expansion (TE) are similar, both being slower than no phase refinement (NR). In real space, the random-omit-map strategy (RO) was slightly faster than simple peak picking (PP) because fewer atoms were used in the structure-factor calculations. Both of these procedures were much faster than iterative peaklist optimization (PO). The original *SHELXD* algorithm (TE + PO) performs quite well in comparison with the *SnB* algorithm (PS + PP) in terms of the percentage of correct solutions, but less well when the efficiency is compared in terms of CPU time per solution. Surprising, the two strategies involving random omit maps (PS + RO and TE + RO), which had been calculated to give reference curves, are much more effective than the other algorithms, especially in terms of CPU efficiency. Indeed these two runs appear to approach a 100% success rate as the number of cycles becomes large. The combination of random omit maps and Karle-type tangent expansion appears to be even more effective (Fig. 16.1.8.4) for gramicidin A, a $P2_12_12_1$ structure (Langs, 1988). It should be noted that conventional direct methods incorporating the tangent formula tend to perform better in $P2_12_12_1$ than in $P1$, perhaps because there is less risk of a uranium-atom pseudosolution.

Subsequent tests using *SHELXD* on several other structures have shown that the use of random omit maps is much more effective than picking the same final number of peaks from the top of the peak list. However, it should be stressed that it is the combination TE + RO that is particularly effective. A possible special case is when a very small number of atoms is sought (*e.g.* Se atoms from MAD data). Preliminary tests indicate that peaklist optimization (PO) is competitive in such cases because the CPU time penalty associated with it is much smaller than when many atoms are involved.



(*a*)

(*b*)

Fig. 16.1.8.3. (*a*) Success rates and (*b*) cost effectiveness for several dual-space strategies as applied to a 148-atom $P1$ structure. The *phase-refinement strategies* are: (PS) parameter-shift reduction of the minimal-function value, (TE) Karle-type tangent expansion (holding the top 40% highest E_c fixed) and (NR) no phase refinement but Sim (1959) weights applied in the E map (these depend on E_c and so cannot be employed after phase refinement). The *real-space strategies* are: (PP) simple peak picking using $0.8N_u$ peaks, (PO) peaklist optimization (reducing N_u peaks to $2N_u/3$), and (RO) random omit maps (also reducing N_u peaks to $2N_u/3$). A total of about 10 000 trials of 400 internal loop cycles each were used to construct this diagram.

Table 16.1.8.4. *Improving success rates by 'completing' the vancomycin data*

Error-free reflections added	Success rate (%)
0	0.25
100 (3.5 Å)	0.3
200 (2.8 Å)	2.1
200 (0.97 Å)	2.4
400 (1.3 Å)	8.2
800 (1.1 Å)	11.1

With hindsight, it is possible to understand why the random omit maps provide such an efficient *search algorithm.* In macromolecular structure refinement, it is standard practice to omit parts of the model that do not fit the current electron density well, to perform some refinement or simulated annealing (Hodel *et al.*, 1992) on the rest of the model to reduce memory effects, and then to calculate a new weighted electron-density map (omit map). If the original features reappear in the new density, they were probably correct; in other cases the omit map may enable a new and better

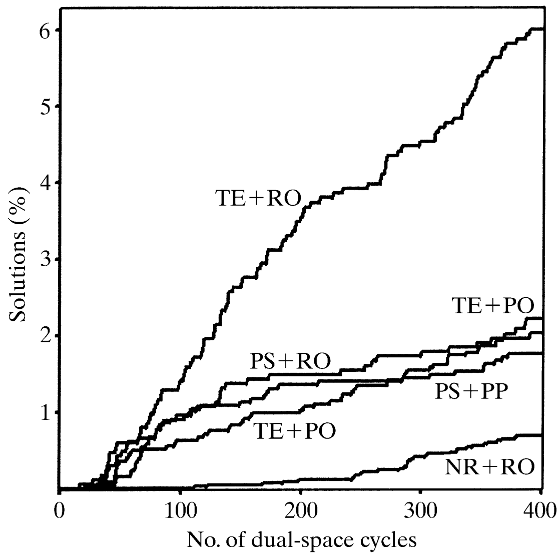


Fig. 16.1.8.4. Success rates for the 317-atom $P2_12_12_1$ structure of gramicidin A.

interpretation. Thus, random omit maps should not lead to the loss of an essentially correct solution, but enable efficient searching in other cases. It is also interesting to note that the results presented in Figs. 16.1.8.3 and 16.1.8.4 show that it is possible, albeit much less efficiently, to solve both structures using random omit maps without the use of any phase relationships based on probability theory (curves NR + RO).

16.1.8.5. *Expansion to P1*

The results shown in Table 16.1.8.4 and Fig. 16.1.8.3 indicate that success rates in space group $P1$ can be anomalously high. This suggests that it might be advantageous to expand all structures to $P1$ and then to locate the symmetry elements afterwards. However, this is more computationally expensive than performing the whole procedure in the true space group, and in practice such a strategy is only competitive in low-symmetry space groups such as $P2_1$, $C2$ or $P\bar{1}$ (Chang *et al.*, 1997). Expansion to $P1$ also offers some opportunities for starting from 'slightly better than random' phases. One possibility, successfully demonstrated by Sheldrick & Gould (1995), is to use a rotation search for a small fragment (*e.g.* a short piece of α-helix) to generate many sets of starting phases; after expansion to $P1$ the translational search usually required for molecular replacement is not needed. Various Patterson superposition minimum functions (Sheldrick & Gould, 1995; Pavelčík, 1994) can also provide an excellent start for phase determination for data expanded to $P1$. Drendel *et al.* (1995) were successful in solving small organic structures *ab initio* by a Fourier recycling method using data expanded to $P1$ without the use of probability theory.

16.1.8.6. *Substructure applications*

It has been known for some time that conventional direct methods can be a valuable tool for locating the positions of heavy-atom substructures using isomorphous (Wilson, 1978) and anomalous (Mukherjee *et al.*, 1989) difference structure factors. Experience has shown that successful substructure applications are highly dependent on the accuracy of the difference magnitudes. As the technology for producing selenomethionine-substituted proteins and collecting accurate multiple-wavelength (MAD) data has improved (Hendrickson & Ogata, 1997; Smith, 1998), there has been an increased need to locate many selenium sites. For larger structures (*e.g.* more than about 30 Se atoms), automated Patterson interpretation methods can be expected to run into difficulties since the number of unique peaks to be analysed increases with the square of the number of atoms. Experimentally measured difference data are an approximation to the data for the hypothetical substructure, and it is reasonable to expect that conventional direct methods might run into difficulties sooner when applied to such data. Dual-space direct methods provide a more robust foundation for handling such data, which are often extremely noisy. Dual-space methods also have the added advantage that the expected number of Se atoms, N_u, which is usually known, can be exploited directly by picking the top N_u peaks. Successful applications require great care in data processing, especially if the F_A values resulting from a MAD experiment are to be used.

All successful applications of *SnB* to previously unknown SeMet data sets, as reported in Table 16.1.8.1, actually involved the use of peak-wavelength anomalous difference data ($|E_\Delta|$). The amount of data available for substructure problems is much larger than for full-structure problems with a comparable number of atoms to be located. Consequently, the user can afford to be stringent in eliminating data with uncertain measurements. Guidelines for rejecting uncertain data have been suggested (Smith *et al.*, 1998). Consideration should be limited to those data pairs ($|E_1|, |E_2|$) [*i.e.*, isomorphous pairs ($|E_{\rm nat}|, |E_{\rm der}|$) and anomalous pairs ($|E_{+\mathbf{H}}|, |E_{-\mathbf{H}}|$)] for which

$$\min[|E_1|/\sigma(|E_1|), |E_2|/\sigma(|E_2|)] \geq x_{\min} \qquad (16.1.8.2)$$

and

$$\frac{||E_1| - |E_2||}{[\sigma^2(|E_1|) + \sigma^2(|E_2|)]^{1/2}} \geq y_{\min}, \qquad (16.1.8.3)$$

where typically $x_{\min} = 3$ and $y_{\min} = 1$. The final choice of maximum resolution to be used should be based on inspection of the spherical shell averages $\langle |E_\Delta|^2 \rangle_s$ *versus* $\langle s \rangle$. The purpose of this precaution is to avoid spuriously large $|E_\Delta|$ values for high-resolution data pairs measured with large uncertainties due to imperfect isomorphism or general fall-off of scattering intensity with increasing scattering angle. Only those $|E_\Delta|$ for which

$$|E_\Delta|/\sigma(|E_\Delta|) \geq z_{\min} \qquad (16.1.8.4)$$

(typically $z_{\min} = 3$) should be deemed sufficiently reliable for subsequent phasing. The probability of very large difference $|E|$'s (*e.g.* > 5) is remote, and data sets that appear to have many such measurements should be examined critically for measurement errors. If a few such data remain even after the adoption of rigorous rejection criteria, it may be best to eliminate them individually. A later paper (Blessing & Smith, 1999) elaborates further data-selection criteria.

On the other hand, it is also important that the phase:invariant ratio be maintained at 1:10 in order to ensure that the phases are overdetermined. Since the largest $|E|$'s for the substructure cell are more widely separated than they are in a true small-molecule cell, the relative number of possible triplets involving the largest reciprocal-lattice vectors may turn out to be too small. Consequently, a relatively small number of substructure phases (*e.g.* $10N_u$) may not have a sufficient number (*i.e.*, $100N_u$) of invariants. Since the number of triplets increases rapidly with the number of reflections considered, the appropriate action in such cases is to increase the number of reflections as suggested in Table 16.1.7.1. This will typically produce the desired overdetermination.

It is rare for Se atoms to be closer to each other than 5 Å, and the application of *SnB* to AdoHcy data truncated to 4 and 5 Å has been successful. Success rates were less for lower-resolution data, but the CPU time required per trial was also reduced, primarily because much smaller Fourier grids were necessary. Consequently, there was no net increase in the CPU time needed to find a solution.

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.

References

16.1

Anderson, D. H., Weiss, M. S. & Eisenberg, D. (1996). *A challenging case for protein crystal structure determination: the mating pheromone Er-1 from Euplotes raikovi. Acta Cryst.* D**52**, 469–480.

Aree, T., Usón, I., Schulz, B., Reck, G., Hoier, H., Sheldrick, G. M. & Saenger, W. (1999). *Variation of a theme: crystal structure with four octakis(2,3,6-tri-O-methyl)-gamma-cyclodextrin molecules hydrated differently by a total of 19.3 water. J. Am. Chem. Soc.* **121**, 3321–3327.

Baggio, R., Woolfson, M. M., Declercq, J.-P. & Germain, G. (1978). *On the application of phase relationships to complex structures. XVI. A random approach to structure determination. Acta Cryst.* A**34**, 883–892.

Beurskens, P. T. (1981). *A statistical interpretation of rotation and translation functions in reciprocal space. Acta Cryst.* A**37**, 426–430.

Bhuiya, A. K. & Stanley, E. (1963). *The refinement of atomic parameters by direct calculation of the minimum residual. Acta Cryst.* **16**, 981–984.

Blessing, R. H. (1997). *LOCSCL: a program to statistically optimize local scaling of single-isomorphous-replacement and single-wavelength-anomalous-scattering data. J. Appl. Cryst.* **30**, 176–177.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Statistical expectation value of the Debye–Waller factor and E(hkl) values for macromolecular crystals. Acta Cryst.* D**52**, 257–266.

Blessing, R. H. & Smith, G. D. (1999). *Difference structure-factor normalization for heavy-atom or anomalous-scattering substructure determinations. J. Appl. Cryst.* **32**, 664–670.

Bricogne, G. (1998). *Bayesian statistical viewpoint on structure determination: basic concepts and examples. Methods Enzymol.* **276**, 361–423.

Burla, M. C., Camalli, M., Cascarano, G., Giacovazzo, C., Polidori, G., Spagna, R. & Viterbo, D. (1989). *SIR88 – a direct-methods program for the automatic solution of crystal structures. J. Appl. Cryst.* **22**, 389–393.

Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. A. (1997). *Incorporating tangent refinement in the Shake-and-Bake formalism. Acta Cryst.* A**53**, 436–444.

Cochran, W. (1955). *Relations between the phases of structure factors. Acta Cryst.* **8**, 473–478.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *Can anomalous signal of sulfur become a tool for solving protein crystal structures? J. Mol. Biol.* **289**, 83–92.

Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Refinement of rubredoxin from Desulfovibrio vulgaris at 1.0 Å with and without restraints. Acta Cryst.* B**48**, 42–59.

Deacon, A. M. & Ealick, S. E. (1999). *Selenium-based MAD phasing: setting the sites on larger structures. Structure*, **7**, R161–R166.

Deacon, A. M., Weeks, C. M., Miller, R. & Ealick, S. E. (1998). *The Shake-and-Bake structure determination of triclinic lysozyme. Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.

Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1985). *On the application of phase relationships to complex structures. XXIV. The Sayre tangent formula. Acta Cryst.* A**41**, 286–290.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *On the application of phase relationships to complex structures. XXII. Techniques for random phase refinement. Acta Cryst.* A**39**, 193–196.

Debaerdemaeker, T. & Woolfson, M. M. (1989). *On the application of phase relationships to complex structures. XXVIII. XMY as a random approach to the phase problem. Acta Cryst.* A**45**, 349–353.

DeTitta, G. T., Edmonds, J. W., Langs, D. A. & Hauptman, H. (1975). *Use of the negative quartet cosine invariants as a phasing figure of merit: NQEST. Acta Cryst.* A**31**, 472–479.

DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Structure solution by minimal-function phase refinement and Fourier filtering. I. Theoretical basis. Acta Cryst.* A**50**, 203–210.

Drendel, W. B., Dave, R. D. & Jain, S. (1995). *Forced coalescence phasing: a method for ab initio determination of crystallographic phases. Proc. Natl Acad. Sci. USA*, **92**, 547–551.

Drouin, M. (1998). Personal communication.

Ekstrom, J. L., Mathews, I. I., Stanley, B. A., Pegg, A. E. & Ealick, S. E. (1999). *The crystal structure of human S-adenosylmethionine decarboxylase at 2.25 Å resolution reveals a novel fold. Structure*, **7**, 583–595.

Fan, H.-F. & Gu, Y.-X. (1985). *Combining direct methods with isomorphous replacement or anomalous scattering data. III. The incorporation of partial structure information. Acta Cryst.* A**41**, 280–284.

Fan, H.-F., Han, F.-S. & Qian, J.-Z. (1984). *Combining direct methods with isomorphous replacement or anomalous scattering data. II. The treatment of errors. Acta Cryst.* A**40**, 495–498.

Fan, H.-F., Hao, Q., Gu, Y.-X., Qian, J.-Z., Zheng, C.-D. & Ke, H. (1990). *Combining direct methods with isomorphous replacement or anomalous scattering data. VII. Ab initio phasing of one-wavelength anomalous scattering data from a small protein. Acta Cryst.* A**46**, 935–939.

Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *A direct-methods solution to the phase problem in the single isomorphous replacement case: theoretical basis and initial applications. Acta Cryst.* A**41**, 571–577.

Frazão, C., Sieker, L., Sheldrick, G. M., Lamzin, V., LeGall, J. & Carrondo, M. A. (1999). *Ab initio structure solution of a dimeric cytochrome c3 from Desulfovibrio gigas containing disulfide bridges. J. Biol. Inorg. Chem.* **4**, 162–165.

Fujinaga, M. & Read, R. J. (1987). *Experiences with a new translation-function program. J. Appl. Cryst.* **20**, 517–521.

Germain, G., Main, P. & Woolfson, M. M. (1970). *On the application of phase relationships to complex structures. II. Getting a good start. Acta Cryst.* B**26**, 274–285.

Germain, G. & Woolfson, M. M. (1968). *On the application of phase relationships to complex structures. Acta Cryst.* B**24**, 91–96.

Gessler, K., Usón, I., Takaha, T., Krauss, N., Smith, S. M., Okada, S., Sheldrick, G. M. & Saenger, W. (1999). *V-Amylose at atomic resolution: X-ray structure of a cycloamylose with 26 glucoses. Proc. Natl Acad. Sci. USA*, **96**, 4246–4251.

Giacovazzo, C. (1976). *A probabilistic theory of the cosine invariant* $\cos(\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}} - \varphi_{\mathbf{h+k+l}})$. *Acta Cryst.* A**32**, 91–99.

Giacovazzo, C. (2001). *Direct methods.* In *International tables for crystallography*, Vol. B. *Reciprocal space*, edited by U. Shmueli, pp. 210–234. Dordrecht: Kluwer Academic Publishers.

Giacovazzo, C. & Platas, J. G. (1995). *The ab initio crystal structure solution of proteins by direct methods. IV. The use of the partial structure. Acta Cryst.* A**51**, 398–404.

Giacovazzo, C., Siliqi, D. & Platas, J. G. (1995). *The ab initio crystal structure solution of proteins by direct methods. V. A new normalizing procedure. Acta Cryst.* A**51**, 811–820.

Giacovazzo, C., Siliqi, D., Platas, J. G., Hecht, H.-J., Zanotti, G. & York, B. (1996). *The ab initio crystal structure solution of proteins by direct methods. VI. Complete phasing up to derivative resolution. Acta Cryst.* D**52**, 813–825.

Giacovazzo, C., Siliqi, D. & Ralph, A. (1994). *The ab initio crystal structure solution of proteins by direct methods. I. Feasibility. Acta Cryst.* A**50**, 503–510.

Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *The ab initio crystal structure solution of proteins by direct methods. II. The procedure and its first applications. Acta Cryst.* A**50**, 609–621.

Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *The ab initio crystal structure solution of proteins by direct methods. III. The phase extension process. Acta Cryst.* A**51**, 177–188.

Hauptman, H. (1974). *On the theory and estimation of the cosine invariants* $\cos(\varphi_{\mathbf{l}} + \varphi_{\mathbf{m}} + \varphi_{\mathbf{n}} + \varphi_{\mathbf{p}})$. *Acta Cryst.* A**30**, 822–829.

Hauptman, H. (1975). *A new method in the probabilistic theory of the structure invariants. Acta Cryst.* A**31**, 680–687.

Hauptman, H. (1982*a*). *On integrating the techniques of direct methods and isomorphous replacement. I. The theoretical basis. Acta Cryst.* A**38**, 289–294.

16.1 (*cont.*)

Hauptman, H. (1982*b*). *On integrating the techniques of direct methods with anomalous dispersion. I. The theoretical basis. Acta Cryst.* A**38**, 632–641.

Hauptman, H., Fisher, J., Hancock, H. & Norton, D. A. (1969). *Phase determination for the estriol structure. Acta Cryst.* B**25**, 811–814.

Hauptman, H. A. (1991). *A minimal principle in the phase problem.* In *Crystallographic computing 5: from chemistry to biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 324–332. Oxford: International Union of Crystallography and Oxford University Press.

Hauptman, H. A. (1996). *The SAS maximal principle: a new approach to the phase problem. Acta Cryst.* A**52**, 490–496.

Hauptman, H. A. & Karle, J. (1953). *Solution of the phase problem. I. The centrosymmetric crystal.* Am. Crystallogr. Assoc. Monograph No. 3. Dayton, Ohio: Polycrystal Book Service.

Hauptman, H. A., Xu, H., Weeks, C. M. & Miller, R. (1999). *Exponential Shake-and-Bake: theoretical basis and applications. Acta Cryst.* A**55**, 891–900.

Hendrickson, W. A. & Ogata, C. M. (1997). *Phase determination from multiwavelength anomalous diffraction measurements. Methods Enzymol.* **276**, 494–523.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Model bias in macromolecular crystal structures. Acta Cryst.* A**48**, 851–858.

Hu, N.-H. & Liu, Y.-S. (1997). *General expression for probabilistic estimation of multiphase structure invariants in the case of a native protein and multiple derivatives. Application to estimates of the three-phase structure invariants. Acta Cryst.* A**53**, 161–167.

Karle, I. L., Flippen-Anderson, J. L., Uma, K., Balaram, H. & Balaram, P. (1989). *α-Helix and mixed $3_{10}/\alpha$-helix in cocrystallized conformers of Boc-Aib-Val-Aib-Aib-Val-Val-Val-Aib-Val-Aib-Ome. Proc. Natl Acad. Sci. USA*, **86**, 765–769.

Karle, J. (1968). *Partial structural information combined with the tangent formula for noncentrosymmetric crystals. Acta Cryst.* B**24**, 182–186.

Karle, J. & Hauptman, H. (1956). *A theory of phase determination for the four types of non-centrosymmetric space groups $1P222$, $2P22$, $3P_12$, $3P_22$. Acta Cryst.* **9**, 635–651.

Kinneging, A. J. & de Graaf, R. A. G. (1984). *On the automatic extension of incomplete models by iterative Fourier calculation. J. Appl. Cryst.* **17**, 364–366.

Kyriakidis, C. E., Peschar, R. & Schenk, H. (1996). *The estimation of four-phase structure invariants using the single difference of isomorphous structure factors. Acta Cryst.* A**52**, 77–87.

Lamzin, V. S. & Wilson, K. S. (1993). *Automatic refinement of protein models. Acta Cryst.* D**49**, 129–147.

Langs, D. A. (1988). *Three-dimensional structure at 0.86 Å of the uncomplexed form of the transmembrane ion channel peptide gramicidin A. Science*, **241**, 188–191.

Langs, D. A. (1993). *Frequency statistical method for evaluating cosine invariants of three-phase relationships. Acta Cryst.* A**49**, 545–557.

Langs, D. A., Guo, D.-Y. & Hauptman, H. A. (1995). *TDSIR phasing: direct use of phase-invariant distributions in macromolecular crystallography. Acta Cryst.* A**51**, 535–542.

Li, C., Kappock, T. J., Stubbe, J., Weaver, T. M. & Ealick, S. E. (1999). *X-ray crystal structure of aminoimidazole ribonucleotide synthetase (PurM), from the Escherichia coli purine biosynthetic pathway at 2.5 Å resolution. Structure*, **7**, 1155–1166.

Loll, P. J., Bevivino, A. E., Korty, B. D. & Axelsen, P. H. (1997). *Simultaneous recognition of a carboxylate-containing ligand and an intramolecular surrogate ligand in the crystal structure of an asymmetric vancomycin dimer. J. Am. Chem. Soc.* **119**, 1516–1522.

Loll, P. J., Miller, R., Weeks, C. M. & Axelsen, P. H. (1998). *A ligand-mediated dimerization mode for vancomycin. Chem. Biol.* **5**, 293–298.

McCourt, M. P., Ashraf, K., Miller, R., Weeks, C. M., Li, N., Pangborn, W. A. & Dorset, D. L. (1997). *X-ray crystal structures of cytotoxic oxidized cholesterols: 7-ketocholesterol and 25-hydroxycholesterol. J. Lipid Res.* **38**, 1014–1021.

McCourt, M. P., Li, N., Pangborn, W., Miller, R., Weeks, C. M. & Dorset, D. L. (1996). *Crystallography of linear molecule binary solids. X-ray structure of a cholesteryl myristate/cholesteryl pentadecanoate solid solution. J. Phys. Chem.* **100**, 9842–9847.

Main, P. (1976). *Recent developments in the MULTAN system – the use of molecular structure.* In *Crystallographic computing techniques*, edited by F. R. Ahmed, pp. 97–105. Copenhagen: Munksgaard.

Main, P., Fiske, S. J., Hull, S. E., Lessinger, L., Germain, G., Declercq, J.-P. & Woolfson, M. M. (1980). *MULTAN80: a system of computer programs for the automatic solution of crystal structures from X-ray diffraction data*. Universities of York, England, and Louvain, Belgium.

Mathiesen, R. H. & Mo, F. (1997). *Application of known triplet phases in the crystallographic study of bovine pancreatic trypsin inhibitor. I: studies at 1.55 and 1.75 Å resolution. Acta Cryst.* D**53**, 262–268.

Mathiesen, R. H. & Mo, F. (1998). *Application of known triplet phases in the crystallographic study of bovine pancreatic trypsin inhibitor. II: study at 2.0 Å resolution. Acta Cryst.* D**54**, 237–242.

Matthews, B. W. & Czerwinski, E. W. (1975). *Local scaling: a method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. Acta Cryst.* A**31**, 480–497.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *On the application of the minimal principle to solve unknown structures. Science*, **259**, 1430–1433.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *SnB: crystal structure determination via Shake-and-Bake. J. Appl. Cryst.* **27**, 613–621.

Mukherjee, A. K., Helliwell, J. R. & Main, P. (1989). *The use of MULTAN to locate the positions of anomalous scatterers. Acta Cryst.* A**45**, 715–718.

Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Ab initio solution and refinement of two high potential iron protein structures at atomic resolution. Acta Cryst.* D**55**, 1773–1784.

Pavelčík, F. (1994). *Patterson-oriented automatic structure determination. Deconvolution techniques in space group P1. Acta Cryst.* A**50**, 467–474.

Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. Acta Cryst.* D**53**, 448–455.

Privé, G. G., Anderson, D. H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). *Packed protein bilayers in the 0.9 Å resolution structure of a designed alpha helical bundle. Protein Sci.* **8**, 1400–1409.

Radfar, R., Shin, R., Sheldrick, G. M., Minor, W., Lovell, C. R., Odom, J. D., Dunlap, R. B. & Lebioda, L. (2000). *The crystal structure of N10-formyltetrahydrofolate synthetase from Moorella thermoacetica. Biochemistry*, **39**, 3920–3926.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Refaat, L. S. & Woolfson, M. M. (1993). *Direct-space methods in phase extension and phase determination. II. Developments of low-density elimination. Acta Cryst.* D**49**, 367–371.

Reibenspiess, J. (1998). Personal communication.

Schäfer, M. (1998). Personal communication.

Schäfer, M. & Prange, T. (1998). Personal communication.

Schäfer, M., Schneider, T. R. & Sheldrick, G. M. (1996). *Crystal structure of vancomycin. Structure*, **4**, 1509–1515.

Schäfer, M., Sheldrick, G. M., Bahner, I. & Lackner, H. (1998). *Crystal structures of actinomycin D and Z3. Angew. Chem.* **37**, 2381–2384.

Schäfer, M., Sheldrick, G. M., Schneider, T. R. & Vértesy, L. (1998). *Structure of balhimycin and its complex with solvent molecules. Acta Cryst.* D**54**, 175–183.

Schenk, H. (1974). *On the use of negative quartets. Acta Cryst.* A**30**, 477–481.

Schneider, T. R. (1998). Personal communication.

16.1 (*cont.*)

Schneider, T. R., Kärcher, J., Pohl, E., Lubini, P. & Sheldrick, G. M. (2000). *Ab initio structure determination of the lantibiotic mersacidin. Acta Cryst.* D**56**, 705–713.

Sha, B.-D., Liu, S.-P., Gu, Y.-X., Fan, H.-F., Ke, H., Yao, J.-X. & Woolfson, M. M. (1995). *Direct phasing of one-wavelength anomalous-scattering data of the protein core streptavidin. Acta Cryst.* D**51**, 342–346.

Sheldrick, G. M. (1982). *Crystallographic algorithms for mini- and maxi-computers*. In *computational crystallography*, edited by D. Sayre, pp. 506–514. Oxford: Clarendon Press.

Sheldrick, G. M. (1990). *Phase annealing in SHELX-90: direct methods for larger structures. Acta Cryst.* A**46**, 467–473.

Sheldrick, G. M. (1997). *Direct methods based on real/reciprocal space iteration*. In *Proceedings of the CCP4 study weekend. Recent advances in phasing*, edited by K. S. Wilson, G. Davies, A. S. Ashton, & S. Bailey, pp. 147–158. DL-CONF-97-001. Warrington: Daresbury Laboratory.

Sheldrick, G. M. (1998). *SHELX: applications to macromolecules*. In *Direct methods for solving macromolecular structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer Academic Publishers.

Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *The application of direct methods and Patterson interpretation to high-resolution native protein data. Acta Cryst.* D**49**, 18–23.

Sheldrick, G. M. & Gould, R. O. (1995). *Structure solution by iterative peaklist optimization and tangent expansion in space group P1. Acta Cryst.* B**51**, 423–431.

Shen, Q. (1998). *Solving the phase problem using reference-beam X-ray diffraction. Phy. Rev. Lett.* **80**, 3268–3271.

Shiono, M. & Woolfson, M. M. (1992). *Direct-space methods in phase extension and phase determination. I. Low-density elimination. Acta Cryst.* A**48**, 451–456.

Shmueli, U. & Wilson, A. J. C. (2001). *Statistical properties of the weighted reciprocal lattice*. In *International tables for crystallography*, Vol. B. *Reciprocal space*, edited by U. Shmueli, pp. 190–209. Dordrecht: Kluwer Academic Publishers.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetical structures. Acta Cryst.* **12**, 813–815.

Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Ab initio structure determination and refinement of a scorpion protein toxin. Acta Cryst.* D**53**, 551–557.

Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. (1998). *The use of SnB to determine an anomalous scattering substructure. Acta Cryst.* D**54**, 799–804.

Smith, J. L. (1998). *Multiwavelength anomalous diffraction in macromolecular crystallography*. In *Direct methods for solving macromolecular structures*, edited by S. Fortier, pp. 211–225. Dordrecht: Kluwer Academic Publishers.

Stec, B., Zhou, R. & Teeter, M. M. (1995). *Full-matrix refinement of the protein crambin at 0.83 Å and 130 K. Acta Cryst.* D**51**, 663–681.

Teichert, M. (1998). Personal communication.

Turner, M. A., Yuan, C.-S., Borchardt, R. T., Hershfield, M. S., Smith, G. D. & Howell, P. L. (1998). *Structure determination of selenomethionyl S-adenosylhomocysteine hydrolase using data at a single wavelength. Nature Struct. Biol.* **5**, 369–375.

Usón, I., Sheldrick, G. M., de La Fortelle, E., Bricogne, G., di Marco, S., Priestle, J. P., Grütter, M. G. & Mittl, P. R. E. (1999). *The 1.2 Å crystal structure of hirustasin reveals the intrinsic flexibility of a family of highly disulphide bridged inhibitors. Structure*, **7**, 55–63.

Vermin, W. J. & de Graaff, R. A. G. (1978). *The use of Karle–Hauptman determinants in small-structure determinations. Acta Cryst.* A**34**, 892–894.

Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). *Refinement of triclinic hen egg-white lysozyme at atomic resolution. Acta Cryst.* D**54**, 522–546.

Wang, B.-C. (1985). *Solvent flattening. Methods Enzymol.* **115**, 90–112.

Weckert, E. & Hümmer, K. (1997). *Multiple-beam X-ray diffraction for physical determination of reflection phases and its applications. Acta Cryst.* A**53**, 108–143.

Weckert, E., Schwegle, W. & Hümmer, K. (1993). *Direct phasing of macromolecular structures by three-beam diffraction. Proc. R. Soc. Lond. Ser. A*, **442**, 33–46.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Structure solution by minimal-function phase refinement and Fourier filtering. II. Implementation and applications. Acta Cryst.* A**50**, 210–220.

Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Applications of the minimal principle to peptide structures. Acta Cryst.* D**49**, 179–181.

Weeks, C. M., Hauptman, H. A., Chang, C.-S. & Miller, R. (1994). *Structure determination by Shake-and-Bake with tangent refinement. ACA Trans. Symp.* **30**, 153–161.

Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Crambin: a direct solution for a 400-atom structure. Acta Cryst.* D**51**, 33–38.

Weeks, C. M. & Miller, R. (1999*a*). *The design and implementation of SnB version 2.0. J. Appl. Cryst.* **32**, 120–124.

Weeks, C. M. & Miller, R. (1999*b*). *Optimizing Shake-and-Bake for proteins. Acta Cryst.* D**55**, 492–500.

Weeks, C. M., Miller, R. & Hauptman, H. A. (1998). *Extending the resolving power of Shake-and-Bake*. In *Direct methods for solving macromolecular structures*, edited by S. Fortier, pp. 463–468. Dordrecht: Kluwer Academic Publishers.

White, P. S. & Woolfson, M. M. (1975). *The application of phase relationships to complex structures. VII. Magic integers. Acta Cryst.* A**31**, 53–56.

Wilson, K. S. (1978). *The application of MULTAN to the analysis of isomorphous derivatives in protein crystallography. Acta Cryst.* B**34**, 1599–1608.

Yao, J.-X. (1981). *On the application of phase relationships to complex structures. XVIII. RANTAN – random MULTAN. Acta Cryst.* A**37**, 642–644.

Zheng, X.-F., Fan, H.-F., Hao, Q., Dodd, F. E. & Hasnain, S. S. (1996). *Direct method structure determination of the native azurin II protein using one-wavelength anomalous scattering data. Acta Cryst.* D**52**, 937–941.

16.2

Bertaut, E. F. (1955*a*). *La méthode statistique en cristallographie. I. Acta Cryst.* **8**, 537–543.

Bertaut, E. F. (1955*b*). *La méthode statistique en cristallographie. II. Quelques applications. Acta Cryst.* **8**, 544–548.

Bricogne, G. (1984). *Maximum entropy and the foundations of direct methods. Acta Cryst.* A**40**, 410–445.

Bricogne, G. (2001). *Fourier transforms in crystallography: theory, algorithms and applications*. In *International tables for crystallography*, Vol. B. *Reciprocal space*, edited by U. Shmueli, 2nd ed., pp. 25–98. Dordrecht: Kluwer Academic Publishers.

Hauptman, H. & Karle, J. (1953). *The solution of the phase problem: I. The centrosymmetric crystal.* ACA Monograph No. 3. Pittsburgh: Polycrystal Book Service.

Jaynes, E. T. (1957). *Information theory and statistical mechanics. Phys. Rev.* **106**, 620–630.

Jaynes, E. T. (1968). *Prior probabilities. IEEE Trans. SSC*, **4**, 227–241.

Jaynes, E. T. (1983). *Papers on probability, statistics and statistical physics.* Dordrecht: Reidel.

Klug, A. (1958). *Joint probability distribution of structure factors and the phase problem. Acta Cryst.* **11**, 515–543.

Shannon, C. E. & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Tsoucaris, G. (1970). *A new method of phase determination. The 'maximum determinant rule'. Acta Cryst.* A**26**, 492–499.

Wiener, N. (1949). *Extrapolation, interpolation and smoothing of stationary time series.* Cambridge, MA: MIT Press.