International Tables for Crystallography (2012). Vol. F, Chapter 16.1, pp. 413–432.

PART 16. DIRECT METHODS

Chapter 16.1. *Ab initio* phasing

G. M. Sheldrick, C. J. Gilmore, 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 α-helices, disulfide bridges or solvent regions) can, however, be utilized. For the last four 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. The direct solution of new macromolecular structures in this way has, however, been limited to a few special cases involving relatively small macromolecules, unusually high-resolution data and, often, the presence of heavier atoms [which might also have been suitable for single-wavelength anomalous diffraction (SAD) or multiple-wavelength anomalous diffraction (MAD) phasing]. However, the same procedures can be applied at much lower resolution for the location of heavy-atom substructures, an essential step in the experimental phasing of macromolecules in the widely used SAD, single isomorphous replacement including anomalous scattering (SIRAS), multiple isomorphous replacement (MIR) and MAD methods. Indeed, substructure-based phasing now accounts for most direct-methods applications to macromolecules. Since three closely related dual-space direct methods computer programs (*SnB*, *SHELXD* and *HySS*) are currently used in the large majority of such applications, we will concentrate on this approach and then describe more briefly some other promising approaches, including one that does not require high-resolution data, a related molecule as search fragment or heavier atoms and should, therefore, be applicable to at least a quarter of the protein structures in the Protein Data Bank (PDB).

16.1.1.1. Data resolution

Direct methods of crystal structure determination have wholly transformed small-molecule crystallography in the past two decades. The same cannot be said for macromolecular crystallography, although there have been very significant advances in the area of substructure determination. The reasons for the success with small molecules are:

(1) automatic and easy to use software is readily and freely available [*e.g. SHELXS* and *SHELXD* (Sheldrick, 1990, 2008; Usón & Sheldrick, 1999), *SnB* (Miller *et al.*, 1994; Weeks & Miller, 1999*a*), *SIR2004* (Burla *et al.*, 2005), and *SUPERFLIP* (Palatinus & Chapuis, 2007)];
(2) the high quality and, in particular, high resolution of data now collected from both laboratory sources and synchrotron facilities; and
(3) data sets are complete with few missing reflections.

Why do data resolution and data quality matter? To understand this, we need to examine a rule proposed by Sheldrick (1990): *Experience with a large number of structures has led us to formulate the empirical rule that, if fewer than half the number of theoretically measurable reflections in the range 1.1 to 1.2 Å are 'observed' [i.e. have $F > 4\sigma(F)$], it is very unlikely that the structure can be solved by direct methods. This critical ratio may be reduced somewhat for centrosymmetric structures and structures containing heavy atoms.*

Morris and Bricogne (2003) offer valuable structural insights into this rule that are instructive for this chapter. By examining the averaged squared normalized structure-factor amplitudes of more than 700 high-resolution (<2.0 Å) structures as a function of data resolution, they found that there is always a pronounced maximum around 1.1 Å, a smaller one around 2.1 Å, and a further pronounced one at ~4.5 Å as shown in Fig. 16.1.1.1. The shape of these curves can be related back to a sinc function transformation which links the intensities of normalized structure-factor profiles and the radial pair distribution function. The peak at 1.1 Å can be shown to arise from bonded distances of ~1.5 Å and non-bonded distances of ~2.4 Å; every protein can be shown to contain distance beats of 1.1 Å arising from these. The net result is a systematic reduction in the expectation value of $|E|^2$ to about 1.25 Å and only then does it rise again. At 1.1–1.2 Å, the resolution is sufficient to reproduce a radial distance distribution with suitably separated peaks, and this gives not only atomicity, but also the stereochemical regularities necessary for the successful application of direct methods. To exacerbate matters further, the fundamental equations of direct methods have variances with a $1/N^{1/2}$ dependence, where N is the number of atoms in the unit cell. Morris and Bricogne make the matter clear: direct methods in their current formulation will always struggle with macromolecular data. This said, however, there are two significant and general uses of direct methods in macromolecular crystallography:

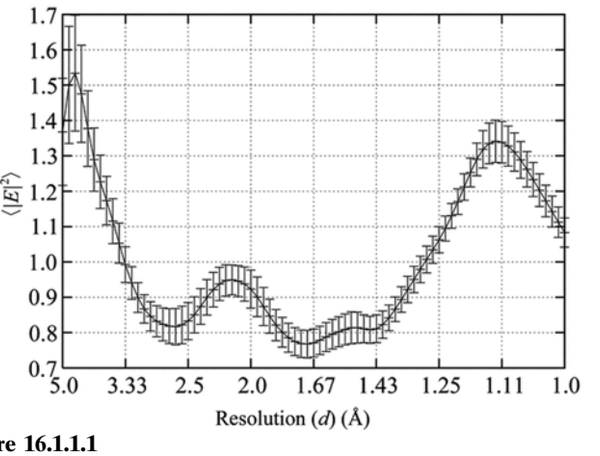


Figure 16.1.1.1
Averaged squared normalized structure-factor amplitudes over 700 protein structures with standard deviations calculated from the population of individual $|E|^2$ profiles (from Morris & Bricogne, 2003).

Copyright © 2012 International Union of Crystallography

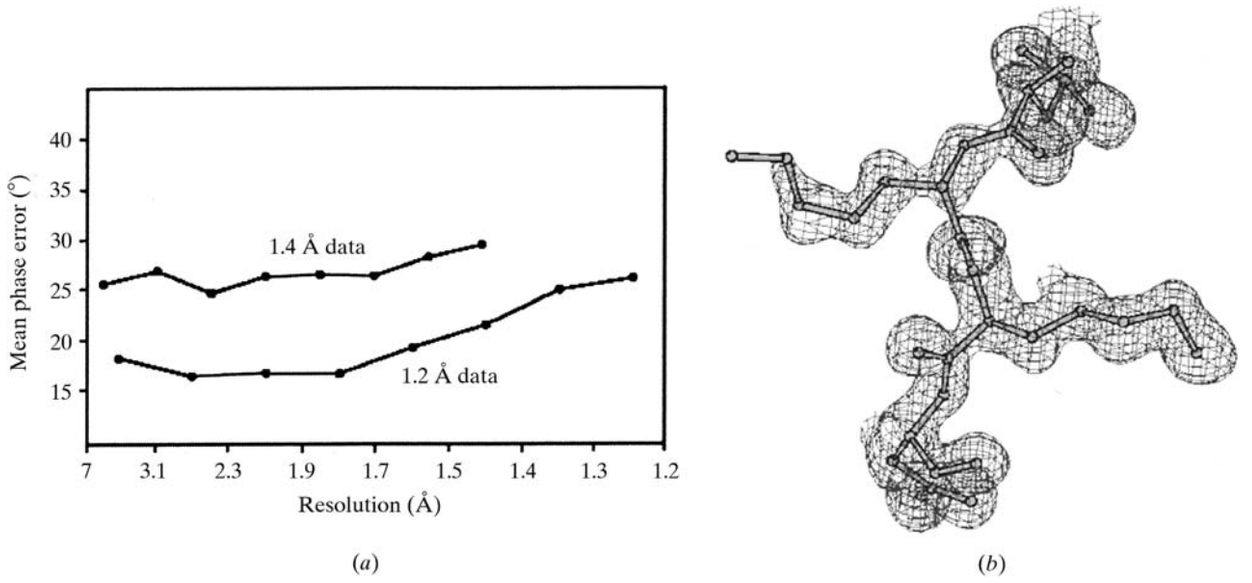


Figure 16.1.1.2
(*a*) Mean phase error as a function of resolution for the two independent *ab initio SHELXD* solutions of the previously unsolved protein hirustasin. Either the 1.2 Å or the 1.4 Å native data set led to solution of the structure. (*b*) Part of the hirustasin molecule from the 1.4 Å room-temperature data after one round of *B*-value refinement with fixed coordinates.

(1) *Ab initio* structure solution with atomic resolution data: the whole crystal structure solution is required with most of the atomic sites sufficiently defined for least-squares refinement. Sheldrick's rule applies rigorously here with very few exceptions in the literature. There have been isolated successes at lower resolutions, but these mostly involved the presence of heavier atoms or data with truncated resolution from crystals that would have diffracted further.
(2) Substructure solution: the determination of the positions of the heavy atoms only (often Se from selenomethionine, but also quite frequently heavy-atom salts, complexes and clusters). Sheldrick's rule is substantially relaxed and structure solutions at data resolutions of 5–6 Å are possible. This follows from the arguments of Morris and Bricogne: the distance beats of 1.1 Å are irrelevant to the substructure. By focusing on the heavy atoms, the complexity of the structure is much reduced, the distance between atoms is larger and the solution of the phase problem becomes easier.

The importance of the presence of several atoms heavier than oxygen for increasing the chance of obtaining a solution by the program *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, which contains ten sulfur and 457 carbon, nitrogen and oxygen atoms in the asymmetric unit, 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. 16.1.1.2(*a*), the mean phase error (MPE) is significantly better for the 1.2 Å data over the whole resolution range. 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. 16.1.1.2*b*), *SHELXD* produced easily interpreted protein density even when bonded atoms are barely resolved from each other.

16.1.1.2. Data completeness

The relative effects of accuracy, completeness and resolution on *Shake-and-Bake* success rates using *SnB* for three large *P*1 structures were studied by computing error-free data using the known atomic coordinates (Xu *et al.*, 2000). The results of these studies, presented in Table 16.1.1.1, 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. As a result of problems with overloaded reflections, the experimental vancomycin data did not include any data at 10 Å resolution or lower. A total of 4000 reflections were phased 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 16.1.1.2 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 synchrotron data for direct methods, it pays to take a second pass using a shorter exposure time to fill in the low-resolution data*.

16.1.1.3. Summary

The basic theory underlying direct methods has been summarized in an excellent chapter (Giacovazzo, 2008) in *International Tables for Crystallography* Volume B (Chapter 2.2) to which the reader is referred for details. Suffice it for this chapter to say that classical direct methods attempt to reconstruct the missing phase information using native data alone by utilizing direct relationships between the crystallographic phases without any *a priori* structural information.

From a historical perspective, the first successful applications of direct methods to native data for structures that could legiti-

Table 16.1.1.1
Success rates for three *P*1 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	—

References: vancomycin: Loll *et al.* (1998); alpha-1: Privé *et al.* (1999); lysozyme: Deacon *et al.* (1998).

Table 16.1.1.2
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

mately be regarded as small macromolecules came from the *Shake-and-Bake* method and the associated *SnB* software (Weeks *et al.*, 1993). 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'). The first previously unknown structures determined by *Shake-and-Bake* were two forms of the 100-atom peptide ternatin (Miller *et al.*, 1993) and, so far, the largest previously unsolved structure solved by direct methods with no atom heavier than oxygen is probably feglymycin, with 1026 unique non-hydrogen atoms and data to 1.10 Å resolution (Bunkóczi *et al.*, 2005).

Using direct methods and accurately measured data, it is now possible to solve heavy-atom substructures of well over 100 atoms. For a state-of-the-art example, see von Delft *et al.* (2003), where a substructure of 160 Se atoms was solved in the product-bound *E. coli* KPHMT using *SnB*. A total of 120 sites were correctly located, allowing the remainder to be located by *SHARP* (de La Fortelle & Bricogne, 1997); in later tests, *SHELXD* was able to find 152 of the sites. For a review of the phase problem in the context of other developments, the reader is referred to a general overview by Dauter (2006).

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), SAD and MAD] to locate substructures at resolutions typically in the range 2.0–3.5 Å, although lower-resolution data are sometimes adequate. 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, this approach might prove relevant in the future. Similarly, the combination of direct methods with multiple-beam diffraction might also play a role (Weckert *et al.*, 1993).

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

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),

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

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

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, 2008). 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 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 SAD 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 SAD magnitudes (Matthews & Czerwinski, 1975; Blessing, 1997), and *DIFFE* for computing the actual difference magnitudes (Blessing & Smith, 1999). The *SnB* program (see Section 16.1.12.4) provides a convenient interface to the *DREAR* suite.

16.1.2.1. SIR differences

Given the individual normalized structure-factor magnitudes for native structures ($|E_{\rm nat}|$) and for structures containing heavy atoms ($|E_{\rm der}|$), as well as 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

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

where $q = q_0\exp(q_1 s^2 + q_2 s^4)$ is a least-squares-fitted empirical renormalization scaling function, dependent on $\sin(\theta)/\lambda$, 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. SAD differences

Given Friedel pairs of normalized structure-factor magnitudes ($|E_{+\mathbf{H}}|$, $|E_{-\mathbf{H}}|$) and the atomic scattering factors (f^0, f_j' and f_j''), then the greatest-lower-bound estimates of SAD difference $|E|$'s are

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

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

In the case of MAD data, the anomalous differences for all wavelengths can be combined in the form of F_A structure factors as described in Chapter 14.2 of this volume by Smith & Hendrickson. In the *SHELX* system, the program *SHELXC* prepares the $|F_A|$ values and the phase shifts α for SAD, SIR, SIRAS, RIP (radiation-damage-induced phasing) and MAD experimental phasing (Sheldrick, 2008, 2010), and the corresponding $|E_A|$ values are derived by *SHELXD*.

16.1.2.3. Difference intensities and direct methods

There are, of course, difficulties with normalized difference intensities. Direct methods normalize the data to give difference E magnitudes, and the macromolecular temperature factors incorporated in this process can be a source of error. The temperature-factor correction, often large for macromolecules, can take small, statistically dubious, amplitude differences at high resolution when working with anomalous difference data and magnify them into large normalized structure-factor differences that are then used in direct methods. These methods are very sensitive to such errors and will often fail unless a suitable resolution truncation limit is imposed. Usually this is around 0.5 Å larger than the experimental diffraction limit of the data or up to the resolution shell with $I/\sigma(I) = 10$. Morris *et al.* (2004) have extended their work on profiles to devise a method of normalization based on $\langle|E|^2\rangle$ profiles, and this should be more robust than traditional methods.

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, there exist 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.

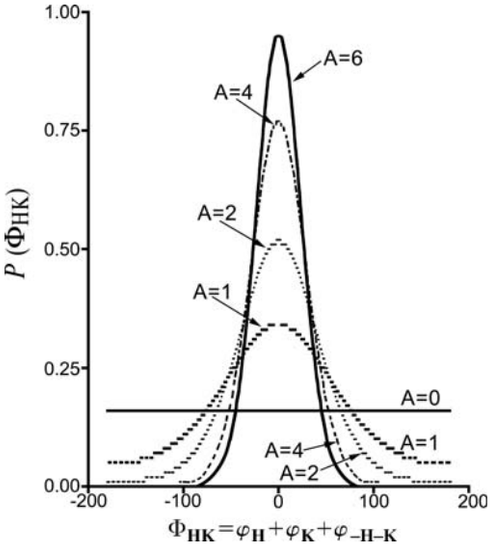


Figure 16.1.3.1
The conditional probability distribution, $P(\Phi_{\mathbf{HK}})$, of the three-phase structure invariants, $\Phi_{\mathbf{HK}}$, having associated parameters $A_{\mathbf{HK}}$ with values of 0, 1, 2, 4 and 6. When A $\simeq$ 0, all values of $\Phi_{\mathbf{HK}}$ are equally likely, and no information useful for phase determination is available. However, the sum of the three phases for most invariants with A $\simeq$ 6 is close to 0°, and an estimate of one phase can be made if the other two are known.

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*,

$$\Phi_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \quad (16.1.3.1)$$

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

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

where

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

and N is the number of atoms, here presumed to be identical, in the 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 corresponding primitive unit cell, N, is small. This is one important 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 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 or polar-axis direction. However, except for the simplest structures, these reflections provide an inadequate foundation for further phase development. Historically, a 'multisolution' or multi-trial approach (Germain & Woolfson, 1968) was 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 occurred, were 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. In addition, some dual-space programs [*SHELXD* (Schneider & Sheldrick, 2002), *HySS* (Grosse-Kunstleve & Adams, 2003)] can also use the Patterson function to generate starting atoms or phases (see Section 16.1.8).

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 three purely reciprocal-space methods have been shown to be of practical value: tangent refinement, parameter-shift optimization of the minimal function, and maximization of Karle–Hauptman determinants (Section 16.1.12.7).

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; Usón & Sheldrick, 1999). 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*,

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

(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. Examples include an exponential expression that has been found to give superior results for several $P1$ structures (Hauptman *et al.*, 1999). In addition, substructure determination using a very simple and computationally efficient modified minimal function,

$$m(\varphi) = 1 - (N_I/N_T) \qquad (16.1.4.3)$$

(where I is an arbitrary interval $[-r, r]$, N_I is the number of triplets whose values lie in I and N_T is the total number of triplets), has been reported (Xu & Hauptman, 2004, 2006; Xu *et al.*, 2005) and incorporated into the *BnP* software (see Section 16.1.12.4).

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*)

For several decades, classical direct methods operated exclusively in reciprocal space, determining phases through statistical relationships between them. Only when this process had converged did the method move into real space by calculating one or more electron-density maps that were examined using stereochemical criteria. In macromolecular crystallography, density modification has always played a central role in phasing. A major advance in direct methods for macromolecules (and large molecules in general) occurred when density-modification methods were incorporated and adapted into the phasing procedure. They are often very simple: peaks which give rise to unrealistic geometries or which are too weak are removed, new structure factors are calculated and hence new phase angles are derived in an iterative process. (They can also be quite sophisticated as in *ACORN2*, which we will discuss in Section 16.1.12.1.) A consequence of this is that the once-clear dividing line between direct methods and other structure-solution techniques has become somewhat blurred.

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 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 within the dual-space direct-methods framework to replace peak picking by alternative density-modification procedures such as low-density elimination (Shiono & Woolfson, 1992; Refaat & Woolfson, 1993) or solvent flattening (Wang, 1985). 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 study by 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, provided that care is taken with disulfide bridges (Section 16.1.11).

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 has been 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 and so is less effective than the random-omit method described in the next section.

16.1.5.3. Random omit maps

A third peak-picking strategy involves selecting approximately $1.3N_u$ 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 the hope of finding an improved position for the deleted fragment, this version of peak picking is described as *random omit*. This procedure helps to prevent the dual-space recycling from getting stuck in a local minimum and is thus an efficient search algorithm.

16.1.6. Fourier refinement

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

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

(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_{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. Resolution enhancement: the 'free lunch' algorithm

Direct methods take a set of phases, refine them and also determine new ones. There is no reason, however, why they cannot be used to predict new amplitudes as well. If density modification of a real-space map is performed, then any process of real-space density modification will, following a Fourier transformation, give structure-factor amplitudes for reflections that were not used to generate it, and these can be outside the resolution limit. If direct methods require atomic resolution data, can we use these techniques to extrapolate structure factors (*i.e.* predict not only phases, but also amplitudes for missing data) and extend data resolution? The idea is not new, but it has been quite extensively studied in recent years. Sheldrick has termed such algorithms 'free lunch', with reference to the saying: 'There is no such thing as a free lunch'! In one example (Usón *et al.*, 2007), weak SIRAS starting phase information followed by density modification led to an $|F_o|$ weighted mean phase error (MPE) of 54° at 1.98 Å resolution, but when the density modification was performed with amplitude extrapolation to 1.0 Å, the MPE fell to 17°. Caliandro *et al.* (2005*a*,*b*) used Patterson or direct methods to obtain trial phases that are submitted to various density-modification methods. Following this, extrapolated phases were generated. This too transformed uninterpretable maps into a solution amenable to automatic tracing. Palatinus *et al.* (2007) used maximum entropy (ME) methods for amplitude extrapolation. In some ways these should be ideal for this purpose, and it is worth noting that ME maps have, *de facto*, optimal resolution enhancement built in, although they can be difficult to generate for large structures.

Why does this work, and why is it sometimes so spectacular? The answer probably lies with the fact that maps are much more sensitive to phases than amplitudes and, if the model bias of predicting new amplitudes is not too great, then using a nonzero value is better than zero, which is the default. Fourier-truncation errors may also be reduced, resulting in less spurious map detail.

16.1.8. Utilizing Pattersons for better starts

When slightly heavier atoms such as sulfur are present, it is possible to start recycling procedures 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. The two-atom vectors are chosen by biased random sampling that favours the vectors corresponding to higher Patterson values. The two atoms

Table 16.1.8.1
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

may be used to generate further atoms using a full Patterson superposition minimum function or a weighted difference synthesis.

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 16.1.8.1 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.

16.1.9. *Shake-and-Bake*: an analysis of a dual-space method in action

The *Shake-and-Bake* algorithm generated the *SnB* program written in Buffalo at the Hauptman–Woodward Institute, principally by Charles Weeks and Russ Miller (Miller *et al.*, 1994; Weeks & Miller, 1999*a*). *SHELXD* (Usón & Sheldrick, 1999; Schneider & Sheldrick, 2002) and later *HySS* (Grosse-Kunstleve & Adams, 2003) were both inspired by *SnB* and employ the *Shake-and-Bake* philosophy with various modifications, in particular involving the use of the Patterson function to obtain starting phases. It is instructive to see how such software works in detail.

16.1.9.1. Flowchart and program comparison

A flowchart for the generic *Shake-and-Bake* algorithm, which provides the foundation for these programs, is presented in Fig. 16.1.9.1. 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. 16.1.9.1, 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* performs Karle-type tangent expansion (Karle, 1968). During tangent or parameter-shift refinement with *SnB*, all phases computed in the preceding structure-factor calculation (steps 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. If there is a tendency for *SHELXD* to produce uranium-atom solutions, more phases should be held fixed in the tangent phase expansion.

(*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. In addition, *SHELXD* calculates a further correlation coefficient, CC_{weak}, calculated in the same

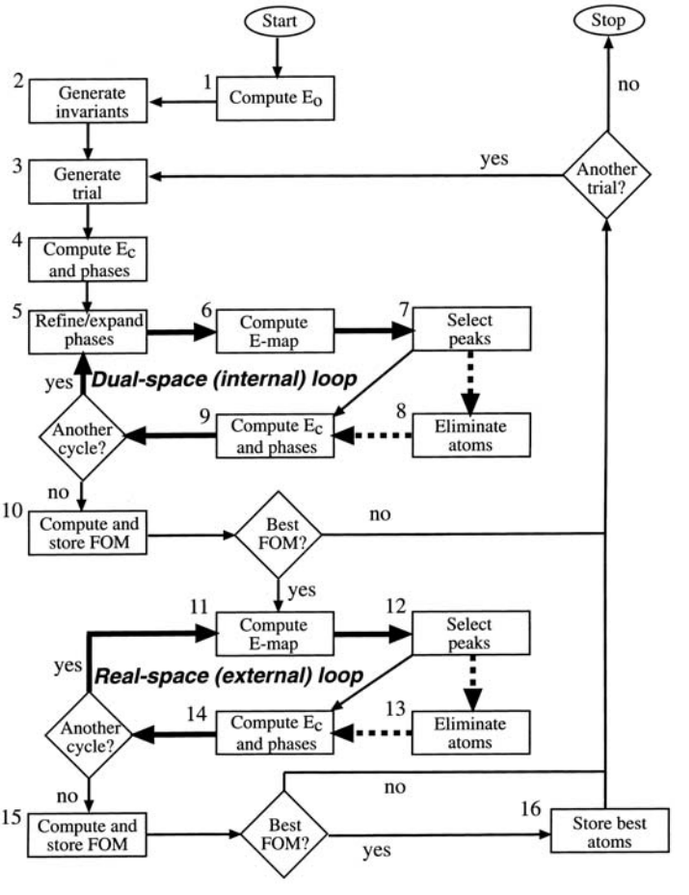


Figure 16.1.9.1
A flowchart for the *Shake-and-Bake* procedure, which is implemented in both *SnB* and *SHELXD*. The essence of the method is the dual-space approach of refining trial structures as they shuttle between real and reciprocal space. In the general case, steps 7 and 12 are any density-modification 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 16.1.9.1
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)

way but using only the weak reflections, *i.e.* those not used directly for phasing.

16.1.9.2. 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 16.1.9.1. 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)] (Deacon *et al.*, 1998). This seems to reduce the frequency of false minima (see Section 16.1.10.1).

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* for *ab initio* solution of the full stucture, and so is used as the default mode (see Section 16.1.10.2 for details). For substructure solution, especially for small substructures, it is normally faster to use starting atoms from Patterson seeding. Although both random omit and Patterson seeding can increase the success rate by an order of magnitude, combining both does not produce much further improvement. For very large substructures, and especially for very high symmetry space groups where the Patterson analysis is more time consuming, the random-omit procedure can be the more effective of the two. The largest substructure solved by *SHELXD* is probably PDB code 2pnk (to be published), solved by Qingping Xu of the Joint Center for Structural Genomics (JCSG), with 197 correct and no incorrect Se sites out of 205 (the other eight were disordered). About 1.6 million trials were needed (using multiple CPUs) to obtain one correct solution when Patterson seeding was employed, but with the random-omit method many good solutions were obtained. This example also illustrates the point that it is important not to give up too soon!

The substructure solution program *HySS* in the *PHENIX* system is more-or-less a clone of *SHELXD*. For further details see Section 16.1.12.5.

16.1.9.3. 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. 16.1.9.2). 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. 16.1.9.3. 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. 16.1.9.3 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

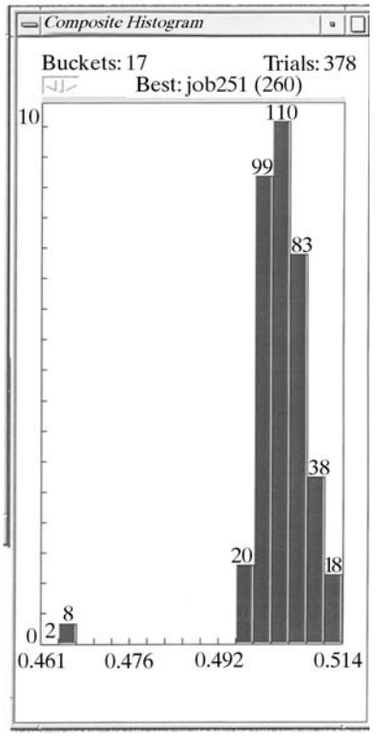


Figure 16.1.9.2
A histogram of figure-of-merit values (minimal function) for 378 scorpion toxin II trials. This bimodal histogram suggests that ten trials are solutions.

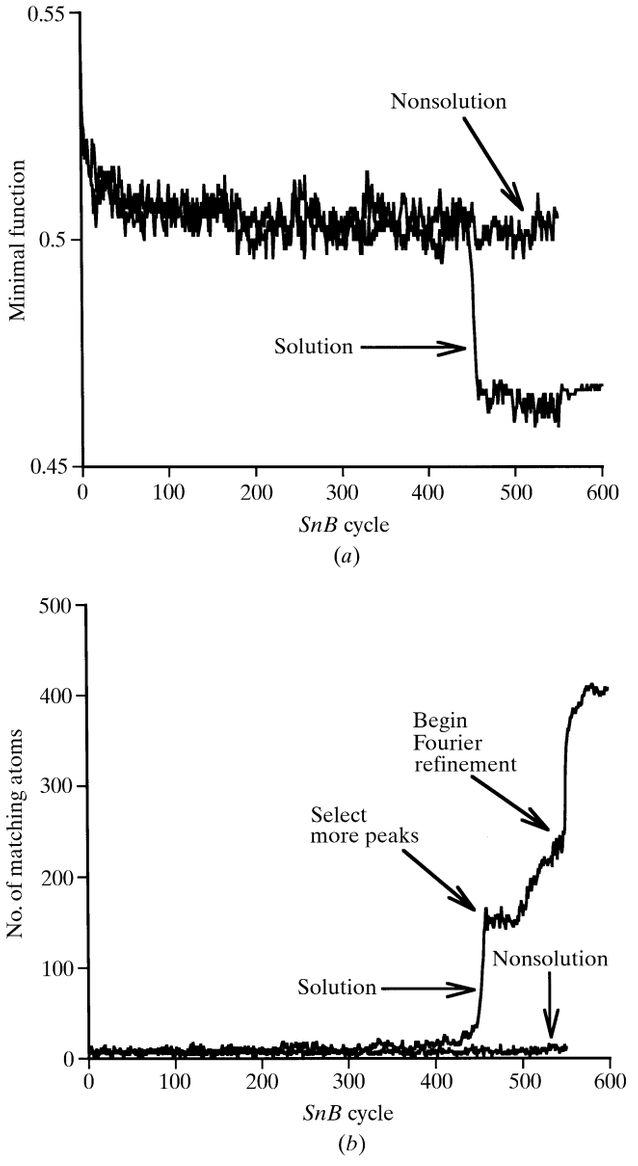


Figure 16.1.9.3
Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function *of Shake-and-Bake* cycle. (*a*) Minimal-function figure of merit, and (*b*) number of peaks closer than 0.5 Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 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.

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, although when the data do not extend to atomic resolution the CC values are less informative, and for a substructure they depend strongly on the data quality.

16.1.10. 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 16.1.10.1. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier-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.

16.1.10.1. 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 *P*1 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 (*C*222) is provided by the Se substructure data for AdoHcy hydrolase (Turner *et al.*, 1998). 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

$$R \text{ ratio} = (R_{\rm con} - R_{\rm unc})/(R_{\rm con} + R_{\rm unc}) \qquad (16.1.10.1)$$

can be used to distinguish false minima from other nonsolutions as well as the true solutions (Xu *et al.*, 2000). 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 *P*1 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.

Table 16.1.10.1
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). The largest substructures solved by these two programs are mentioned in the text.

Compound	Space group	N_u (molecule)	N_u + solvent	N_u (heavy)	Resolution (Å)	Program	Reference
Hirustasin	$P4_32_12$	402	467	10S	1.2–1.55	D	[1]
Cyclodextrin derivative	$P2_1$	448	467	—	0.88	D	[2]
Alpha-1 peptide	$P1$	408	471	Cl	0.92	S	[3]
Rubredoxin*	$P2_1$	395	497	Fe, 6S	1.0–1.1	S, D	[4]
Vancomycin	$P1$	404	547	12Cl	0.97	S	[5]
BPTI*	$P2_12_12_1$	453	561	7S	1.08	D	[6]
Cyclodextrin derivative	$P2_1$	504	562	28S	1.00	D	[7]
Balhimycin*	$P2_1$	408	598	8Cl	0.96	D	[8]
Mg-complex*	$P1$	576	608	8Mg	0.87	D	[9]
Scorpion toxin II*	$P2_12_12_1$	508	624	8S	0.96–1.2	S	[10]
Bucandin	$C2$	516	634	10S	1.05	D	[11]
Decaplanin	$P2_1$	448	635	4Cl	1.00	D	[12]
Amylose-CA26	$P1$	624	771	—	1.10	D	[13]
Viscotoxin B2	$P2_12_12_1$	722	818	12S	1.05	D	[14]
Mersacidin	$P3_2$	750	826	24S	1.04	D	[15]
Cv HiPIP H42Q*	$P2_12_12_1$	631	837	4Fe	0.93	D	[16]
Feglymycin	$P6_5$	828	1026	—	1.10	D	[17]
Acutohaemolysin	$C2_1$	1010	1242	17S	0.8	S	[18]
Tsuchimycin	$P1$	1069	1293	24Ca	1.00	D	[19]
HEW lysozyme*	$P1$	1001	1295	10S	0.85	S, D	[20], [21]
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20	D	[16]
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20	D	[22]

References: [1] Usón *et al.* (1999); [2] Aree *et al.* (1999); [3] Privé *et al.* (1999); [4] Dauter *et al.* (1992); [5] Loll *et al.* (1998); [6] Schneider (1998); [7] Reibenspiess *et al.* (2000); [8] Schäfer *et al.* (1998); [9] Teichert (1998); [10] Smith *et al.* (1997); [11] Kuhn *et al.* (2000); [12] Lehmann *et al.* (2003); [13] Gessler *et al.* (1999); [14] Pal *et al.* (2008); [15] Schneider *et al.* (2000); [16] Parisini *et al.* (1999); [17] Bunkóczi *et al.* (2005); [18] Liu *et al.* (2003); [19] Bunkóczi (2004); [20] Deacon *et al.* (1998); [21] Walsh *et al.* (1998); [22] Frazão *et al.* (1999).

It should be noted that *SHELXD* optionally deletes the highest peak if the second peak is less than a specified fraction (*e.g.* 40%) of the height of the first, in an attempt to 'kick' the structure out of a false minimum.

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. 16.1.10.1), 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 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 (Schäfer *et al.*, 1998) is an example of a large non-*P*1 structure that also requires a parameter shift of around 154° to obtain a solution using the minimal function.

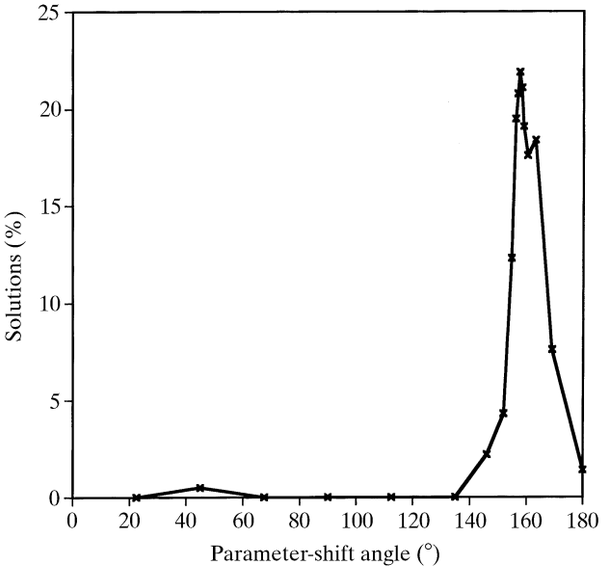
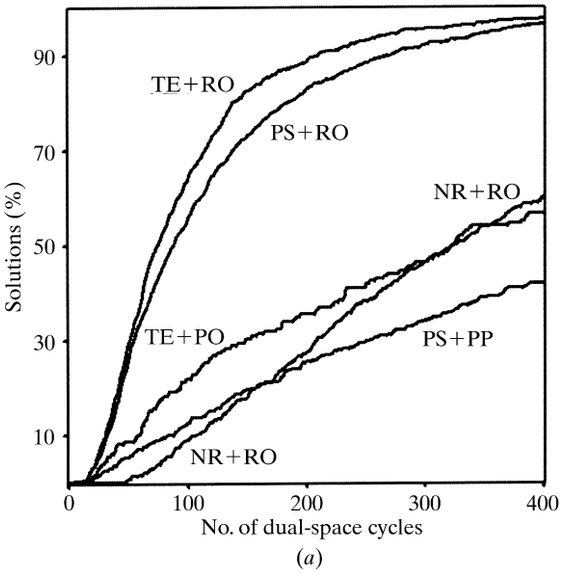


Figure 16.1.10.1
Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

16.1.10.2. 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*. Fig. 16.1.10.2 shows the results of different strategies tested on a 148-atom *P*1 structure (Karle *et al.*, 1989) while developing *SHELXD*. 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 included in the test as placebos, 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



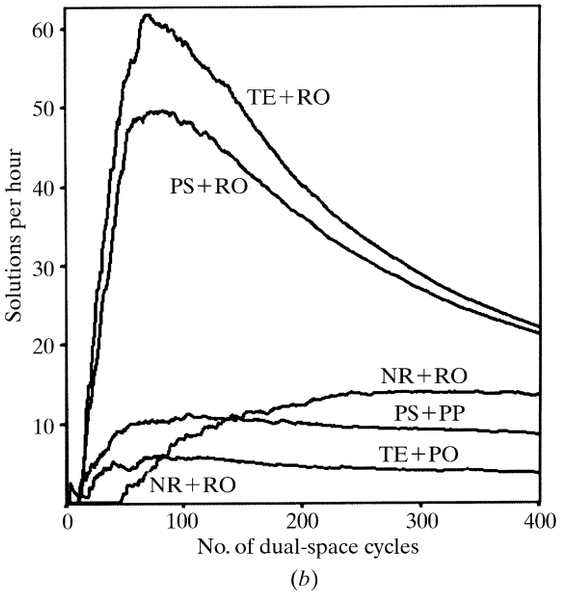


Figure 16.1.10.2
(*a*) Success rates and (*b*) cost effectiveness for several dual-space strategies as applied to a 148-atom *P*1 structure. The *phase-refinement strategies* are: (PS) parameter-shift reduction of the minimal-function value, (TE) Karle-type tangent expansion (holding the top 40% highest 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.

expansion appears to be even more effective (Fig. 16.1.10.3) 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 for this space group than in *P*1, perhaps because there is less risk of a uranium-atom pseudo-solution.

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

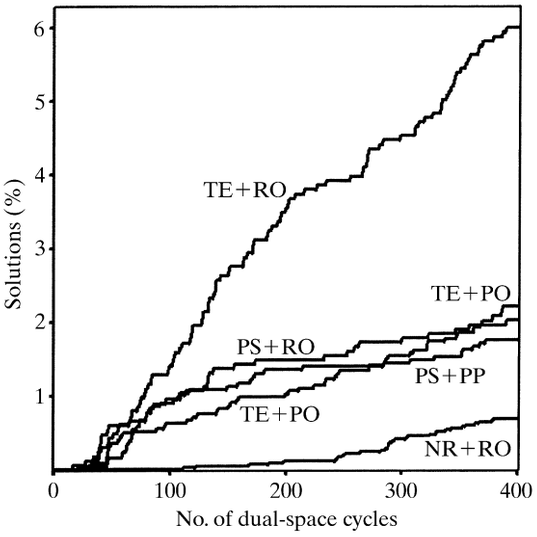


Figure 16.1.10.3
Success rates for the 317-atom $P2_12_12_1$ structure of gramicidin A.

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.

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 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. 16.1.10.2 and 16.1.10.3 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).

16.1.10.3. Expansion to P1

The results shown in Table 16.1.1.1 and Fig. 16.1.10.2 indicate that success rates in space group *P*1 can be anomalously high. This suggests that it might be advantageous to expand all structures to *P*1 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 $P2_1$, *C*2 or $P\bar{1}$ (Chang *et al.*, 1997). Expansion to *P*1 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 *P*1 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 *P*1. Drendel *et al.* (1995) were successful in solving small organic structures *ab initio* by a Fourier recycling method using perturbed Fourier amplitudes and data expanded to *P*1 without the use of

probability theory. The random-omit procedure combined with expansion to *P*1 in *SHELXD* also enables structures to be solved efficiently even when the tangent formula phase extension is switched off; this has the advantage that lower E values can be used than would be suitable for the tangent formula, but at the cost of increasing the CPU time per solution. The program *ACORN2* (Dodson & Woolfson, 2009) is also particularly effective in *P*1; it applies sophisticated density modification and dual-space recycling with a special density disturbance term (POWDM) that is applied every tenth cycle. For the *P*1 form of lysozyme (see Table 16.1.10.1), good phases can be obtained by *ACORN2* starting from a fragment as small as two sulfur atoms for the 0.93 Å data. Expansion of the data to *P*1 is an essential feature of the charge-flipping approach described in Section 16.1.12.6. In general, one can say that dual-space recycling of data expanded to *P*1 requires some reasonable perturbation of the density (*e.g.* charge flipping or random peak omit) to prevent stagnation, but with this precaution provides a simple and effective approach to structure solution.

16.1.10.4. 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.

SHELXD is frequently successfully employed with $|F_A|$ values derived from multiwavelength MAD data generated, for example, by the programs *SHELXC* (Sheldrick, 2008, 2010) or *XPREP* (Bruker AXS, Madison, WI). The decision at which resolution the data should be truncated for substructure determination is best taken on the basis of the correlation coefficients between the signed anomalous differences (Schneider & Sheldrick, 2002). On the other hand, *SnB* is normally applied separately to anomalous and dispersive differences. In many cases, both approaches lead to successful substructure solution. The real advantage of MAD data is that they provide more experimental phase information (*i.e.* better maps) and this is most important at medium to low resolution. 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

$$\min\left[|E_1|/\sigma(|E_1|), |E_2|/\sigma(|E_2|)\right] \geq x_{\rm min} \qquad (16.1.10.2)$$

and

$$\frac{\|E_1| - |E_2\|}{[\sigma^2(|E_1|) + \sigma^2(|E_2|)]^{1/2}} \geq y_{\rm min}, \qquad (16.1.10.3)$$

where typically $x_{\rm min} = 3$ and $y_{\rm min} = 1$. 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* $\langle s\rangle$ where $s = \sin(\theta)/\lambda$. 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 $|E_\Delta|$'s for which

$$|E_\Delta|/\sigma(|E_\Delta|) \geq z_{\rm min} \qquad (16.1.10.4)$$

(typically $z_{\rm min} = 3$) 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 paper by 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 16.1.9.1. 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 hydolase 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.

16.1.11. Substructure solution for native sulfurs and halide soaks

In the past, experimental phasing usually involved either the preparation of selenomethionine derivatives or the incorporation of heavy-metal ions by soaking crystals with a low concentration of the metal salt for several hours. The first of these methods required time in the wet lab and did not work well for all expression systems; the second had a low success rate. The improved quality of modern diffraction data collected from cryo-cooled crystals makes it now possible to exploit the weak anomalous signal from the native sulfur atoms or from halide ions introduced by soaking with a high concentration of a halide (iodide or bromide) for a few seconds immediately before cryocooling the crystal (Dauter *et al.*, 2000, 2001; Usón *et al.*,

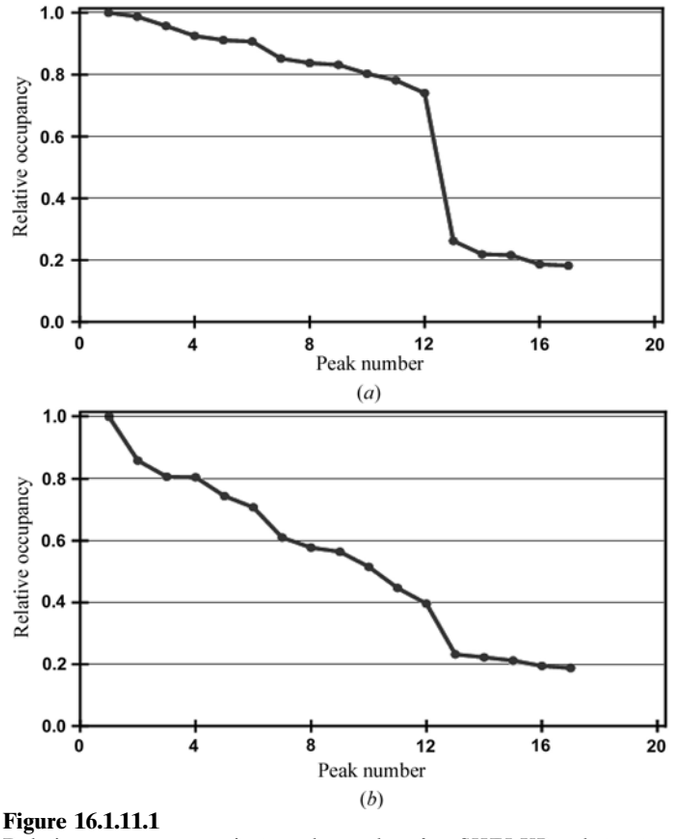


Figure 16.1.11.1
Relative occupancy against peak number for *SHELXD* substructure solutions of elastase. (*a*) Sulfur-SAD experiment showing the presence of the 12 expected sulfur atoms. (*b*) Iodide soak. Subsequent analysis showed that the peaks with relative occupancies less than 0.2 are mainly noise. These figures were made with *HKL2MAP* (Pape & Schneider, 2004).

2003). The success of these approaches is also made possible by the ability of modern, dual-space, substructure-solution programs to locate correctly a large number of sites, possibly with varying occupancies, using the SAD and SIRAS approaches.

In selenomethionine SAD and MAD phasing and in sulfur SAD phasing, the variation of the occupancies (refined in the final two cycles in the case of *SHELXD*) provides a very good indication as to whether the structure has been solved. Fig. 16.1.11.1(*a*) shows the phasing of elastase with sulfur SAD; a sharp drop in the relative occupancy after the 12th site confirms the expected presence of 12 sulfur atoms. For an iodide soak of the same protein (Fig. 16.1.11.1*b*), the relative occupancies show a gradual fall with peak number. Since the number of sites is difficult to estimate in advance for a halide soak and *SHELXD* needs to know this number approximately (within say 20%), it may be necessary to make several trials with different numbers of expected sites. From experience, the best number to use is the one that causes the occupancies to fall to about 0.2 relative to the strongest peak. Usually, subsequent refinements of the occupancies show that all the sites are partially occupied for halide soaks.

When the anomalous signal does not extend beyond about 2.0 Å, the two sulfur atoms of a disulfide bridge coalesce to a single maximum, often referred to as a supersulfur atom. At low resolution, this increases the signal-to-noise ratio for such sites in the dual-space procedure, but tends to impede phase extension to higher resolution (*e.g.* when density modification is applied to the native data with the starting phases estimated using these supersulfur atoms). An efficient way around this problem is to fit dumbbells rather than single atoms in the peak-search part of the dual-space recycling (Debreczeni *et al.*, 2003); this dramatically improves the quality of the higher-resolution starting phases.

Because the weak anomalous signal is swamped by the noise at higher resolution in such SAD experiments, it is often essential to truncate the resolution of the anomalous difference data before searching for the substructure. For MAD experiments, it is customary to truncate the data to the resolution at which the correlation coefficient between the signed anomalous differences falls below 30% (Schneider & Sheldrick, 2002). The same criterion can be used for SAD experiments if two independent data sets (*e.g.* from two different crystals) are available. As a compromise, the signed anomalous differences can be divided randomly into two sets, and then the correlation coefficient between them can be calculated. However, since these sets are not completely independent, a higher threshold (say 40%) might be advisable. An alternative criterion is to truncate the data at the point where the ratio of the mean absolute anomalous difference to its mean standard deviation falls below ~1.3, but this requires rather precise estimates of the standard deviations. In borderline cases, especially when multiple CPUs are available, it is probably safer simply to run the substructure solution for a range of different resolution cutoffs in parallel, and this is already implemented in several of the automated phasing pipelines. Sometimes good solutions are only obtained in a rather limited resolution cutoff range. A good starting value for sulfur SAD is the diffraction limit plus 0.5 Å.

16.1.12. Computer programs for dual-space phasing

Macromolecular crystallography is well served with free, high-quality, open-source software. Programs that provide direct-methods phasing for macromolecular problems will now be outlined. Although they all (except *CRUNCH2*) implement procedures that can be described more-or-less as dual-space methods, there are also appreciable differences from the three programs discussed so far. In this section, we have attempted to highlight these differences.

16.1.12.1. ACORN

ACORN (Yao *et al.*, 2006) and its successor *ACORN2* (Dodson & Woolfson, 2009) start with a fragment. This fragment can be very small: 1–8% in *ACORN*, and as little as 0.25% of the scattering is reported for *ACORN2*. Strictly speaking, these are not direct-methods programs, since they solve and refine crystal structures from poor starting phase sets that are usually derived from a known fragment. However, since this fragment can be very small, and since for *P*1 structures a single heavy atom at the origin suffices as a useable starting point, they are included here.

The data are normalized to give E magnitudes and partitioned into three sets: (1) large observed normalized magnitudes, (2) small magnitudes (typically < 0.2), and (3) the unobserved reflections (which are given values of unity) for a resolution range. A fragment is used to generate a set of phases, and this is followed by a sophisticated density-modification procedure:

$$\begin{aligned}
\rho^{(n+1)} &= 0 \quad \text{if} \quad \rho^{(n)} \leq L\sigma, \\
\rho^{(n+1)} &= \rho^{(n)} \tanh[0.2(\rho^{(n)}/\sigma)^{\eta}] \quad \text{if} \quad \rho^{(n)} > L\sigma, \\
\rho^{(n+1)} &= T\sigma \quad \text{if} \quad \rho^{(n+1)} > T\sigma,
\end{aligned} \tag{16.1.12.1}$$

where σ is the standard deviation of the map density and

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

where M is the number of observable reflections within the resolution sphere and N is the number of atoms in the unit cell (excluding H atoms). The unconstrained value of T_1 is approximately 0.5 of the expected peak height of the heaviest atom in the E map with perfect phases; c is the cycle number.

$$L = L_1 - L_1^{c/n},$$
$$L_1 = 1.05\left[\left(B/r^2\right) - 1\right]\Phi(Z_{\max}), \qquad (16.1.12.3)$$

where B is the usual overall temperature factor; Φ is a cubic function going through the points (Φ, Z) = (0.84, 16), (0.96, 30), (1.15, 34.5) and (1.24, 48). If $Z_{\max} < 16$, the value $Z = 16$ (corresponding to sulfur) is used, and for $Z_{\max} > 48$, the value $Z = 48$ (corresponding to cadmium) is used. The value of L is thus reduced in n cycles from L_1 to zero.

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

where 'nint' indicates the nearest integer and

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

Finally,

$$\eta = 17.24(r-1)^5 + 1.5. \qquad (16.1.12.6)$$

This use of η improves performance at low resolution. There is also an E limit,

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

imposed with the proviso that $0.75 \le E_{\text{lim}} \le 1.25$. Resolution enhancement is simple: all missing data can optionally be given a value of 1.0, which is the square root of the expectation value of E^2 from Wilson statistics. It can be seen that this procedure always needs a fragment, but this can be very small indeed. Examples include five S atoms in the catalytic domain of chitinase A1 from *Bacillus circulans* WL-12 (Matsumoto *et al.*, 1999) that account for less than 1% of the scattering density and oxidoreductase (Haynes *et al.*, 1994) with less than 0.5% of the scattering power in the initial fragment. In the latter case, the space group is $P1$ so one heavy atom can be placed at the unit-cell origin. Two atoms define the fragment: one was placed at the origin and the second in a position compatible with the Patterson map. The final mean phase error was 31.6°. *ACORN2* is available as part of the *CCP4* package at http://www.ccp4.ac.uk.

16.1.12.2. IL MILIONE

The *IL MILIONE* software is a product of the Bari group (Burla *et al.*, 2007). It provides a complete suite of programs for structure solution including software for processing SAD, MAD or SIRAS data or for molecular replacement. We will focus here on the *ab initio* procedures in this package. *Ab initio* phasing uses triplets and tangent procedures or Patterson methods. Direct-space refinement using density modification is employed along with the use of a resolution extension procedure. For the initial phasing, triplet invariants are evaluated by means of the P10 formula (Cascarano *et al.*, 1984). The tangent formula [equation (16.1.4.1)] is used in conjunction with these triplet-phase estimates starting with random phases and multiple starting points. An early figure of merit (eFOM) is calculated for each tangent trial and only the best trial solutions based on this are submitted to direct-space refinement. When misplaced molecular fragments are present, the structures can often be solved by the *RELAX* procedure (Burla *et al.*, 2003). In this procedure, the phases of a trial solution obtained in the correct space group are extended and refined in $P1$ by direct-space techniques. The appropriate figures of merit are used to determine the appropriate vector shift to operate on the fragment.

Patterson deconvolution may also be used in conjunction with direct methods. A superposition minimum function is used. The first peak from this procedure is always used in the phasing process. For each of the remaining peaks (the number of which depends on the size of the structure to be solved and on its data resolution), a set of phases is obtained which is ranked by a specific early figure of merit (pFOM) defined as

$$\text{pFOM} = \sigma/R(\Phi), \qquad (16.1.12.8)$$

where σ is the usual standard deviation of the electron-density map and $R(\Phi)$ is the minimal function [equation (16.1.4.2)]. Irrespective of whether direct or Patterson methods are used, direct-space refinement techniques consist of cycles of electron-density modification (in which only a small fraction of the electron densities are inverted to obtain new phases), and a mask from molecular envelope calculations is applied. The correct solution is identified by a final figure of merit (fFOM), and the program automatically stops when fFOM exceeds a given threshold that depends on data resolution and structural complexity. The Patterson deconvolution methods proved to be, by far, the most efficient ones for large structures and, therefore, have been chosen as the default procedure in this case. They extended the size of the macromolecular structures that are solvable *ab initio* to more than 6000 non-H atoms in the asymmetric unit, provided that at least one calcium atom is present in the asymmetric unit and atomic resolution data are available (Burla *et al.*, 2007).

In favourable circumstances, *IL MILIONE* is also able to solve protein structures with data resolution up to 1.4–1.5 Å and to provide interpretable electron-density maps. The program has been tested using about 100 structures randomly taken from among those in the PDB with resolution better than 1.6 Å. It was able to solve all the test structures that had atomic resolution data, less than 2000 non-H atoms in the asymmetric unit (Nasym), and atoms heavier than calcium present (Zmax > 20). The solution efficiency is reduced to 84% for structures with atomic resolution data, Zmax > 20 and Nasym > 2000, and it is reduced further when these conditions are not fully met. In the presence of atoms as heavy as Ho, Au, Hg or Yb, solutions of structures composed of more than 1000 atoms have been achieved at resolutions as low as 2.0 Å. Finally, it was successful even at 1.65 Å for a case containing up to 7890 non-hydrogen atoms in the asymmetric unit (Caliandro *et al.*, 2008).

IL MILIONE can also apply direct methods to SIR and MIR data. Two different approaches may be followed for protein crystal-structure solution from isomorphous data (up to five derivatives may be used). The triplet phase invariants are estimated *via* the conditional probability distribution function,

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

(Hauptman, 1982*a*; Giacovazzo *et al.*, 1988, 1996), where Φ_p is the triplet phase of the protein and

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

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

16.1.12.3. SHELX

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

16.1.12.4. SnB and BnP

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

16.1.12.5. HySS

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

16.1.12.6. SUPERFLIP: charge flipping

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

16.1.12.7. CRUNCH2 – Karle–Hauptman determinants

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

16.1.13. Conclusions and the grand challenge

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

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

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

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

Flippen-Anderson, Carlos Frazão, Katrin Gessler, Håkon Hope, Jörg Kärcher, Peter Kuhn, Victor Lamzin, David Langs, Christopher Lehmann, 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.

References

Adams, P. W., Afonine, P. V., Grosse-Kunstleve, R. W., Moriarty, N. W., Sauter, N. K., Zwart, P. H., Gopal, K., Ioerger, T. R., Kanbi, L., McKee, E., Pai, R. K., Hung, L.-W., Radhakannan, T., McCoy, A. J., Read, R. J., Storoni, L. C., Romo, T. D., Sacchettini, J. C. & Terwilliger, T. C. (2007). *Automated structure determination with Phenix*. In *Evolving Methods for Macromolecular Crystallography*, edited by R. J. Read & J. L. Sussman. pp. 101–109. Dordrecht: Springer.

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.

Bunkóczi, G. (2004). *Structure determination of peptides with antimicrobial action*. PhD Thesis, Georg-August-Universität, Göttingen, Germany.

Bunkóczi, G., Vértesy, L. & Sheldrick, G. M. (2005). *The antiviral antibiotic feglymycin: First direct-methods solution of a 1000+ equal-atom structure*. *Angew. Chem. Int. Ed.* **44**, 1340–1342.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G., Siliqi, D. & Spagna, R. (2007). *IL MILIONE: a suite of computer programs for crystal structure solution of proteins*. *J. Appl. Cryst.* **40**, 609–613.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Spagna, R. (2005). *SIR2004: an improved tool for crystal structure determination and refinement*. *J. Appl. Cryst.* **38**, 381–388.

Burla, M. C., Carrozzini, B., Caliandro, R., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Polidori, G. (2003). *Ab initio protein phasing at 1.4 Å resolution: the new phasing approach of SIR2003-N*. *Acta Cryst.* A**59**, 560–568.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2008). *Ab initio phasing of proteins with heavy atoms at non-atomic resolution: pushing the size limit of solvable structures up to 7890 non-H atoms in the asymmetric unit*. *J. Appl. Cryst.* **41**, 548–553.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*a*). *Phasing at resolution higher than the experimental resolution*. *Acta Cryst.* D**61**, 556–565.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*b*). *Ab initio phasing at resolution higher than experimental resolution*. *Acta Cryst.* D**61**, 1080–1087.

Cascarano, G., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *The method of representations of structure seminvariants. The strengthening of triplet relationships*. *Acta Cryst.* A**40**, 278–283.

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. (2006). *Current state and prospects of macromolecular crystallography*. *Acta Cryst.* D**62**, 1–11.

Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Novel approach to phasing proteins: derivatization by short cryo-soaking with halides*. *Acta Cryst.* D**56**, 232–237.

Dauter, Z., Li, M. & Wlodawer, A. (2001). *Practical experience with the use of halides for phasing macromolecular structures: a powerful tool for structural genomics*. *Acta Cryst.* D**57**, 239–249.

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., 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. & 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.

Debreczeni, J. É., Girmann, B., Zeeck, A., Krätzner, R. & Sheldrick, G. M. (2003). *Structure of viscotoxin A3: disulfide location from weak SAD data*. *Acta Cryst.* D**59**, 2125–2132.

Delft, F. von, Inoue, T., Saldanha, S. A., Ottenhof, H. H., Schmitzberger, F., Birch, L. M., Dhanaraj, V., Witty, M., Smith, A. G., Blundell, T. L. & Abell, C. (2003). *Structure of E. coli ketopantoate hydroxymethyl transferase complexed with ketopantoate and Mg2+, solved by locating 160 selenomethionine sites*. *Structure*, **11**, 985–996.

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.

Dodson, E. J. & Woolfson, M. M. (2009). *ACORN2: new developments of the ACORN concept*. *Acta Cryst.* D**65**, 881–891.

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.

Dumas, C. & van der Lee, A. (2008). *Macromolecular structure by charge flipping*. *Acta Cryst.* D**64**, 864–873.

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.

Fu, Z.-Q., Chrzas, J., Sheldrick, G. M., Rose, J. & Wang, B.-C. (2007). *A parallel program using SHELXD for quick heavy-atom partial structure solution on high-performance computers*. *J. Appl. Cryst.* **40**, 387–390.

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

Gelder, R. de, de Graaff, R. A. G. & Schenk, H. (1993). *Automatic determination of crystal structures using Karle–Hauptman matrices*. *Acta Cryst.* A**49**, 287–293.

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. (2008). *Direct methods.* In *International Tables for Crystallography*, Vol. B. *Reciprocal Space*, edited by U. Shmueli, ch. 2.2. Dordrecht: Kluwer Academic Publishers.

Giacovazzo, C., Cascarano, G. & Zheng, C. D. (1988). *On integrating the techniques of direct methods and isomorphous replacement.*

A new probabilistic formula for triplet invariants. *Acta Cryst.* A**44**, 45–51.

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.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Substructure search procedures for macromolecular structures*. *Acta Cryst.* D**59**, 1966–1973.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework*. *J. Appl. Cryst.* **35**, 126–136.

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.

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. 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. & 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.

Haynes, M. R., Stura, E. A., Hilvert, D. & Wilson, I. A. (1994). *Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 Å*. *Science*, **263**, 646–652.

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.

Karle, I. L., Flippen-Anderson, J. L., Uma, K., Balaram, H. & Balaram, P. (1989). *α-Helix and mixed 3_{10}/α-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. (1950). *The phases and magnitudes of the structure factors*. *Acta Cryst.* **3**, 181–187.

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.

Kuhn, P., Deacon, A. M., Comoso, S., Rajaseger, G., Kini, R. M., Usón, I. & Kolatkar, P. R. (2000). *The atomic resolution structure of bucandin, a novel toxin isolated from the Malayan krait, determined by direct methods*. *Acta Cryst.* D**56**, 1401–1407.

La Fortelle, E. de & Bricogne, G. (1997). *Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods*. *Methods Enzymol.* **276**, 472–494.

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.

Lehmann, C., Debreczeni, J. É., Bunkcóczi, G., Dauter, M., Dauter, Z., Vértesy, L. & Sheldrick, G. M. (2003). *Structures of four crystal forms of decaplanin*. *Helv. Chim. Acta*, **86**, 1478–1497.

Liu, Q., Huang, Q., Teng, M., Weeks, C. M., Jelsch, C., Zhang, R. & Niu, L. (2003). *The crystal structure of a novel, inactive, lysine 49 PLA2 from Agkistrodon acutus venom: an ultrahigh resolution, ab initio structure determination*. *J. Biol. Chem.* **278**, 41400–41408.

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.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Phaser crystallographic software*. *J. Appl. Cryst.* **40**, 658–674.

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.

Matsumoto, T., Nonata, T., Hashimoto, M., Watanabe, T. & Mitsui, Y. (1999). *Three-dimensional structure of the catalytic domain of chitinase Al from Bacillus circulars WL-12 at a very high resolution*. *Proc. Jpn. Acad. Ser. B*, **75**, 269–274.

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.

Meijers, R., Hans-Werner, A., Dauter, Z., Wilson, K. S., Lamzin, V. S. & Cedergren-Zeppezauer, E. S. (2007). *Structural evidence for a ligand coordination switch in liver alcohol dehydrogenase*. *Biochemistry*, **46**, 5446–5454.

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.

Miller, R., Shah, N., Green, M. L., Furey, W. & Weeks, C. M. (2007). *Shake-and-Bake on the grid*. *J. Appl. Cryst.* **40**, 938–944.

Morris, R. J., Blanc, E. & Bricogne, G. (2004). *On the interpretation and use of $\langle|E|^2\rangle(d^*)$ profiles*. *Acta Cryst.* D**60**, 227–240.

Morris, R. J. & Bricogne, G. (2003). *Sheldrick's 1.2 Å rule and beyond*. *Acta Cryst.* D**59**, 615–617.

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.

Ness, S. R., de Graaff, R. A. G., Abrahams, J. P. & Pannu, N. S. (2004). *Crank: new methods for automated macromolecular crystal structure solution*. *Structure*, **12**, 1753–1761.

Oszlányi, G. & Sütő, A. (2004). *Ab initio structure solution by charge flipping*. *Acta Cryst.* A**60**, 134–141.

Oszlányi, G. & Sütő, A. (2005). *Ab initio structure solution by charge flipping. II. Use of weak reflections*. *Acta Cryst.* A**61**, 147–152.

Oszlányi, G. & Sütő, A. (2008). *The charge flipping algorithm*. *Acta Cryst.* A**64**, 123–134.

Pal, A., Debreczeni, J. É., Sevvana, M., Gruene, T., Kahle, B., Zeeck, A. & Sheldrick, G. M. (2008). *Structures of viscotoxins A1 and B2 from European mistletoe solved using native data alone*. *Acta Cryst.* D**64**, 985–992.

Palatinus, L. & Chapuis, G. (2007). *SUPERFLIP – a computer program for the solution of crystal structures by charge flipping in arbitrary dimensions*. *J. Appl. Cryst.* **40**, 786–790.

Palatinus, L., Steurer, W. & Chapuis, G. (2007). *Extending the charge-flipping method towards structure solution from incomplete data sets*. *J. Appl. Cryst.* **40**, 456–462.

Pape, T. & Schneider, T. R. (2004). *HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs*. *J. Appl. Cryst.* **37**, 843–844.

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.

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.

Rappleye, J., Innus, M., Weeks, C. M. & Miller, R. (2002). *SnB version 2.2: an example of crystallographic multiprocessing*. *J. Appl. Cryst.* **35**, 374–376.

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. H., Maynard, D. K., Derecskei-Kovacs, A. & Vigh, G. (2000). *Crystal structures of heptakis(2,6-di-O-tert-butyldimethylsilyl)-cyclomaltoheptaose, heptakis(2-O-methyl-3,6-di-O-tert-butyldimethyl-silyl)cyclomaltoheptaose and heptakis(2-O-methyl)cyclomalto heptaose*. *Carbohydrate Res.* **328**, 217–227.
Rodriguez, D. D., Grosse, C., Himmel, S., Gonzales, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Crystallographic ab initio protein structure solution below atomic resolution*. *Nat. Methods*, **6**, 651–653.
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.
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.
Schneider, T. R. & Sheldrick, G. M. (2002). *Substructure solution with SHELXD*. *Acta Cryst.* D**58**, 1772–1779.
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. (2002). *Macromolecular phasing with SHELXE*. *Z. Kristallogr.* **217**, 644–650.
Sheldrick, G. M. (2008). *A short history of SHELX*. *Acta Cryst.* A**64**, 112–122.
Sheldrick, G. M. (2010). *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification*. *Acta Cryst.* D**66**, 479–485.
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.
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. (2008). *Statistical properties of the weighted reciprocal lattice*. *International Tables for Crystallography*, Vol. B. *Reciprocal Space*, edited by U. Shmueli, ch. 2.1. 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.
Teichert, M. (1998). *Strukturaufklärung von tetranuklearen Magnesium-chelatkomplexen – Datensammlungsstrategien mit dem SMART-CCD-System*. PhD Thesis, Georg-August-Universität, Göttingen, Germany.
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. Nat. Struct. Biol.* **5**, 369–375.
Usón, I., Schmidt, B., von Bülow, R., Grimme, S., von Figura, K., Dauter, M., Rajashankar, K. R., Dauter, Z. & Sheldrick, G. M. (2003). *Locating the anomalous scatterer substructures in halide and sulfur phasing*. *Acta Cryst.* D**59**, 57–66.
Usón, I. & Sheldrick, G. M. (1999). *Advances in direct methods for protein crystallography*. *Curr. Opin. Struct. Biol.* **9**, 643–648.
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.
Usón, I., Stevenson, C. E. M., Lawson, D. M. & Sheldrick, G. M. (2007). *Structure determination of the O-methyltransferase NovP using the 'free lunch algorithm' as implemented in SHELXE*. *Acta Cryst.* D**63**, 1069–1074.
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., 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., Blessing, R. H., Miller, R., Mungee, R., Potter, S. A., Rappleye, J., Smith, G. D., Xu, H. & Furey, W. (2002). *Towards automated protein structure determination: BnP, the SnB-PHASES interface*. *Z. Kristallogr.* **217**, 686–693.
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. & 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.
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.
Xu, H. & Hauptman, H. A. (2004). *Statistical approach to the phase problem*. *Acta Cryst.* A**60**, 153–157.
Xu, H. & Hauptman, H. A. (2006). *Recent advances in direct phasing for heavy-atom substructure determination*. *Acta Cryst.* D**62**, 897–900.
Xu, H., Weeks, C. M., Deacon, A. M., Miller, R. & Hauptman, H. A. (2000). *Ill-conditioned Shake-and-Bake: the trap of the false minimum*. *Acta Cryst.* A**56**, 112–118.
Xu, H., Weeks, C. M. & Hauptman, H. A. (2005). *Optimizing statistical Shake-and-Bake for Se-atom substructure determination*. *Acta Cryst.* D**61**, 976–981.
Yao, J.-X. (1981). *On the application of phase relationships to complex structures. XVIII. RANTAN – random MULTAN. Acta Cryst.* A**37**, 642–644.
Yao, J. X, Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006). *ACORN: a review*. *Acta Cryst.* D**62**, 901–908.