International Tables for Crystallography (2012). Vol. F, Section 16.1.1, pp. 413–415.

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:

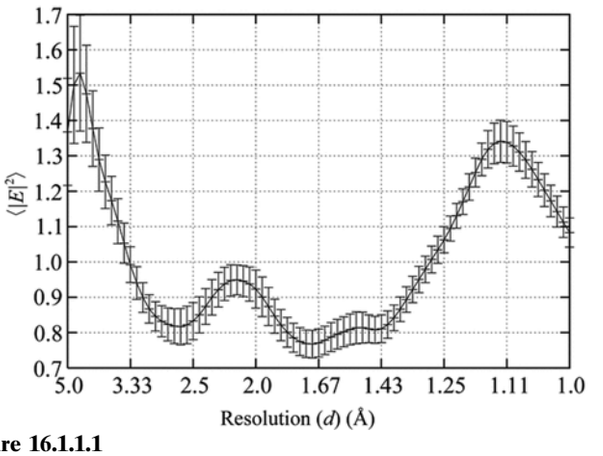

Figure 16.1.1.1
Averaged squared normalized structure-factor amplitudes over 700 protein structures with standard deviations calculated from the population of individual $|E|^2$ profiles (from Morris & Bricogne, 2003).

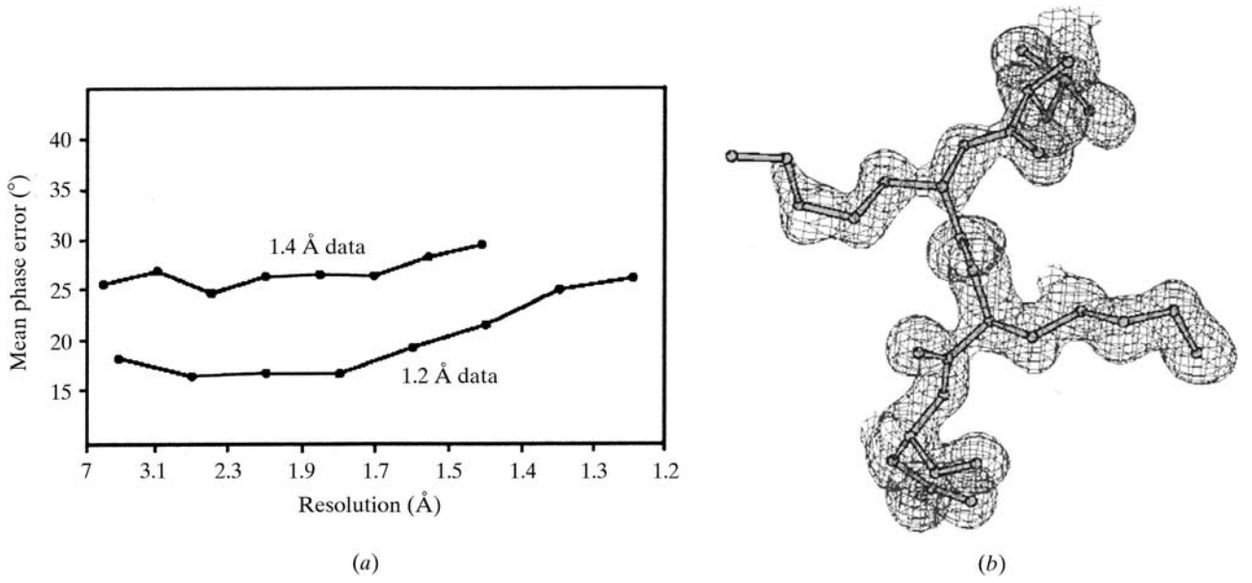

Figure 16.1.1.2
(*a*) Mean phase error as a function of resolution for the two independent *ab initio SHELXD* solutions of the previously unsolved protein hirustasin. Either the 1.2 Å or the 1.4 Å native data set led to solution of the structure. (*b*) Part of the hirustasin molecule from the 1.4 Å room-temperature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

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

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

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.