

## 16.1. AB INITIO PHASING

**Table 16.1.1.1**

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

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

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 feqlymycin, 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

**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

future. Similarly, the combination of direct methods with multiple-beam diffraction might also play a role (Weckert *et al.*, 1993).

**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),

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

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

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,  $B_{\text{iso}}$  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), *LOCSCAL* 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.