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:

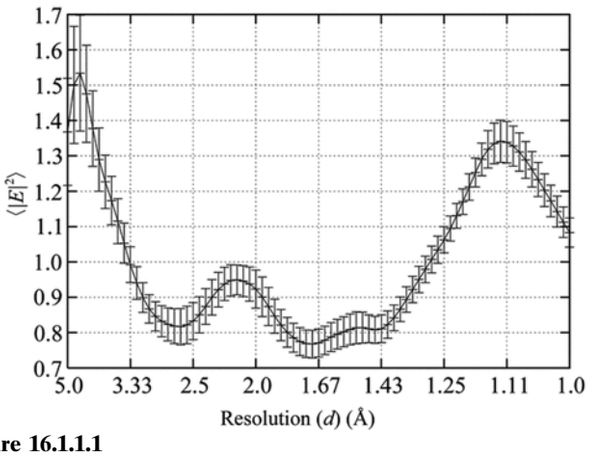

Figure 16.1.1.1
Averaged squared normalized structure-factor amplitudes over 700 protein structures with standard deviations calculated from the population of individual $|E|^2$ profiles (from Morris & Bricogne, 2003).

references