International Tables for Crystallography (2006). Vol. B, Section 2.2.10.2, pp. 231–232.

2.2.10. Direct methods in macromolecular crystallography

2.2.10.1. *Introduction*

Protein structures cannot be solved *ab initio* by traditional direct methods (*i.e.*, by application of the tangent formula alone). Accordingly, the first applications were focused on two tasks:

(*a*) improvement of the accuracy of the available phases (refinement process);

(*b*) extension of phases from lower to higher resolution (phase-extension process).

The application of standard tangent techniques to (*a*) and (*b*) has not been found to be very satisfactory (Coulter & Dewar, 1971; Hendrickson *et al.*, 1973; Weinzierl *et al.*, 1969). Tangent methods, in fact, require atomicity and non-negativity of the electron density. Both these properties are not satisfied if data do not extend to atomic resolution ($d > 2$ Å). Because of series termination and other errors the electron-density map at $d > 2$ Å presents large negative regions which will appear as false peaks in the squared structure. However, tangent methods use only a part of the information given by the Sayre equation (2.2.6.5). In fact, (2.2.6.5) express two equations relating the radial and angular parts of the two sides, so obtaining a large degree of overdetermination of the phases. To achieve this Sayre (1972) [see also Sayre & Toupin (1975)] suggested minimizing (2.2.10.1) by least squares as a function of the phases:

$$\sum_{\mathbf{h}} \left| a_{\mathbf{h}} F_{\mathbf{h}} - \sum_{\mathbf{k}} F_{\mathbf{k}} F_{\mathbf{h}-\mathbf{k}} \right|^2 . \qquad (2.2.10.1)$$

Even if tests on rubredoxin (extensions of phases from 2.5 to 1.5 Å resolution) and insulin (Cutfield *et al.*, 1975) (from 1.9 to 1.5 Å resolution) were successful, the limitations of the method are its high cost and, especially, the higher efficiency of the least-squares method. Equivalent considerations hold for the application of determinantal methods to proteins [see Podjarny *et al.* (1981); de Rango *et al.* (1985) and literature cited therein].

A question now arises: why is the tangent formula unable to solve protein structures? Fan *et al.* (1991) considered the question from a first-principle approach and concluded that:

(1) the triplet phase probability distribution is very flat for proteins (N is very large) and close to the uniform distribution;

(2) low-resolution data create additional problems for direct methods since the number of available phase relationships per reflection is small.

Sheldrick (1990) suggested that direct methods are not expected to succeed if fewer than half of the reflections in the range 1.1–1.2 Å are observed with $|F| > 4\sigma(|F|)$ (a condition seldom satisfied by protein data).

The most complete analysis of the problem has been made by Giacovazzo, Guagliardi *et al.* (1994). They observed that the expected value of α (see Section 2.2.7) suggested by the tangent formula for proteins is comparable with the variance of the α parameter. In other words, for proteins the signal determining the phase is comparable with the noise, and therefore the phase indication is expected to be unreliable.

2.2.10.2. *Ab initio direct phasing of proteins*

Section 2.2.10.1 suggests that the mere use of the tangent formula or the Sayre equation cannot solve *ab initio* protein structures of usual size. However, even in an *ab initio* situation, there is a source of supplementary information which may be used. Good examples are the 'peaklist optimization' procedure (Sheldrick & Gould, 1995) and the *SIR*97 procedure (Altomare *et al.*, 1999) for refining and completing the trial structure offered by the first E map.

In both cases there are reasons to suspect that the correct structure is sometimes extracted from a totally incorrect direct-methods solution. These results suggest that a direct-space procedure can provide some form of structural information complementary to that used in reciprocal space by the tangent or similar formulae. The combination of real- and reciprocal-space techniques could therefore enlarge the size of crystal structures solvable by direct methods. The first program to explicitly propose the combined use of direct and reciprocal space was *Shake and Bake* (*SnB*), which inspired a second package, *half-bake* (*HB*). A third program, *SIR*99, uses a different algorithm.

The *SnB* method (DeTitta *et al.*, 1994; Weeks *et al.*, 1994; Hauptman, 1995) is the heir of the *cosine least-squares method* described in Section 2.2.8, point (4). The function

$$R(\Phi) = \frac{\sum_j G_j [\cos \Phi_j - D_1(G_j)]^2}{\sum_j G_j},$$

where Φ is the triplet phase, $G = 2|E_{\mathbf{h}} E_{\mathbf{k}} E_{\mathbf{h}+\mathbf{k}}|/(N)^{1/2}$ and $D_1(x) = I_1(x)/I_0(x)$.

$R(\Phi)$ is expected to have a global minimum, provided the number of phases involved is sufficiently large, when all the phases are equal to their true values for some choice of origin and enantiomorph. Thus the phasing problem reduces to that of finding the global minimum of $R(\Phi)$ (the *minimum principle*).

SnB comprises a *shake* step (phase refinement) and a *bake* step (electron-density modification), the second step aiming to impose phase constraints implicit in real space. Accordingly, the program requires two Fourier transforms per cycle, and numerous cycles. Thus it may be very time consuming and it is not competitive with other direct methods for the solution of the crystal structures of small molecules. However, it introduced into the field the tremendous usefulness of intensive computations for the direct solution of complex crystal structures.

Owing to Sheldrick (1997), *HB* does most of its work in direct space. Random atomic positions are generated, to which a modified *peaklist optimization* process is applied. A number of peaks are eliminated subject to the condition that $\sum |E_c|(|E_0|^2 - 1)$ remains as large as possible (only reflections with $|E_0| > |E_{\min}|$ are involved, where $|E_{\min}| \simeq 1.4$). The phases of a suitable subset of reflections are then used as input for a tangent expansion. Then an E map is calculated from which peaks are selected: these are submitted to the elimination procedure.

Typically 5–20 cycles of this internal loop are performed. Then a correlation coefficient (CC) between $|E_0|$ and $|E_c|$ is calculated for all the data. If the CC is good (*i.e.* larger than a given threshold), then a new loop is performed: a new E map is obtained, from which a list of peaks is selected for submission to the elimination procedure. The criterion now is the value of the CC, which is calculated for all the reflections. Typically two to five cycles of this external loop are performed.

The program works indefinitely, restarting from random atoms until interrupted. It may work either by applying the true space-group symmetry or after having expanded the data to $P1$.

The *SIR*99 procedure (Burla *et al.*, 1999) may be divided into two distinct parts: the tangent section (*i.e.*, a double tangent process using triplet and quartet invariants) is followed by a real-space refinement procedure. As in *SIR*97, the reciprocal-space part is followed by the real-space refinement, but this time this last part is much more complex. It involves three different techniques: EDM (an electron-density modification process), the HAFR part (in which all the peaks are associated with the heaviest atomic species) and the DLSQ procedure (a least-squares Fourier refinement process). The atomicity is gradually introduced into the procedure. The entire process requires, for each trial, several cycles of EDM and HAFR:

the real-space part is able to lead to the correct solution even when the tangent formula does not provide favourable phase values.

2.2.10.3. *Integration of direct methods with isomorphous replacement techniques*

The modulus of the isomorphous difference

$$\Delta F = |F_{PH}| - |F_P|$$

may be assumed at a first approximation as an estimate of the heavy-atom s.f. F_H. Normalization of $|\Delta F|$'s and application of the tangent formula may reveal the heavy-atom structure (Wilson, 1978).

The theoretical basis for integrating the techniques of direct methods and isomorphous replacement was introduced by Hauptman (1982*a*). According to his notation let us denote by f_j and g_j atomic scattering factors for the atom labelled j in a pair of isomorphous structures, and let $E_{\mathbf{h}}$ and $G_{\mathbf{h}}$ denote corresponding normalized structure factors. Then

$$E_{\mathbf{h}} = |E_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) = \alpha_{20}^{-1/2} \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j),$$

$$G_{\mathbf{h}} = |G_{\mathbf{h}}| \exp(i\psi_{\mathbf{h}}) = \alpha_{02}^{-1/2} \sum_{j=1}^{N} g_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j),$$

where

$$\alpha_{mn} = \sum_{j=1}^{N} f_j^m g_j^n.$$

The conditional probability of the two-phase structure invariant $\Phi = \varphi_{\mathbf{h}} - \psi_{\mathbf{h}}$ given $|E_{\mathbf{h}}|$ and $|G_{\mathbf{h}}|$ is (Hauptman, 1982*a*)

$$P(\Phi|\ |E|, |G|) \simeq [2\pi I_0(Q)]^{-1} \exp(Q \cos \Phi),$$

where

$$Q = |EG|[2\alpha/(1-\alpha^2)],$$
$$\alpha = \alpha_{11}/(\alpha_{20}^{1/2}\alpha_{02}^{1/2}).$$

Three-phase structure invariants were evaluated by considering that eight invariants exist for a given triple of indices $\mathbf{h}$, $\mathbf{k}$, $\mathbf{l}$ ($\mathbf{h}+\mathbf{k}+\mathbf{l}=0$):

$$\Phi_1 = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}} \qquad \Phi_2 = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \psi_{\mathbf{l}}$$
$$\Phi_3 = \varphi_{\mathbf{h}} + \psi_{\mathbf{k}} + \varphi_{\mathbf{l}} \qquad \Phi_4 = \psi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}}$$
$$\Phi_5 = \varphi_{\mathbf{h}} + \psi_{\mathbf{k}} + \psi_{\mathbf{l}} \qquad \Phi_6 = \psi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \psi_{\mathbf{l}}$$
$$\Phi_7 = \psi_{\mathbf{h}} + \psi_{\mathbf{k}} + \varphi_{\mathbf{l}} \qquad \Phi_8 = \psi_{\mathbf{h}} + \psi_{\mathbf{k}} + \psi_{\mathbf{l}}.$$

So, for the estimation of any Φ_j, the joint probability distribution

$$P(E_{\mathbf{h}}, E_{\mathbf{k}}, E_{\mathbf{l}}, G_{\mathbf{h}}, G_{\mathbf{k}}, G_{\mathbf{l}})$$

has to be studied, from which eight conditional probability densities can be obtained:

$$P(\Phi_i \| E_{\mathbf{h}}|, |E_{\mathbf{k}}|, |E_{\mathbf{l}}|, |G_{\mathbf{h}}|, |G_{\mathbf{k}}|, |G_{\mathbf{l}}|)$$
$$\simeq [2\pi I_0(Q_j)]^{-1} \exp[Q_j \cos \Phi_j]$$

for $j = 1, \ldots, 8$.

The analytical expressions of Q_j are too intricate and are not given here (the reader is referred to the original paper). We only say that Q_j may be positive or negative, so that reliable triplet phase estimates near 0 or near π are possible: the larger $|Q_j|$, the more reliable the phase estimate.

A useful interpretation of the formulae in terms of experimental parameters was suggested by Fortier *et al.* (1984): according to them, distributions do not depend, as in the case of the traditional three-phase invariants, on the total number of atoms per unit cell but rather on the scattering difference between the native protein and the derivative (that is, on the scattering of the heavy atoms in the derivative).

Hauptman's formulae were generalized by Giacovazzo *et al.* (1988): the new expressions were able to take into account the resolution effects on distribution parameters. The formulae are completely general and include as special cases native protein and heavy-atom isomorphous derivatives as well as X-ray and neutron diffraction data. Their complicated algebraic forms are easily reduced to a simple expression in the case of a native protein heavy-atom derivative: in particular, the reliability parameter for Φ_1 is

$$Q_1 = 2[\sigma_3/\sigma_2^{3/2}]_P |E_{\mathbf{h}} E_{\mathbf{k}} E_{\mathbf{l}}| + 2[\sigma_3/\sigma_2^{3/2}]_H \Delta_{\mathbf{h}} \Delta_{\mathbf{k}} \Delta_{\mathbf{l}}, \quad (2.2.10.2)$$

where indices P and H warn that parameters have to be calculated over protein atoms and over heavy atoms, respectively, and

$$\Delta = (F_{PH} - F_P)/(\textstyle\sum f_j^2)_H^{1/2}.$$

Δ is a pseudo-normalized difference (with respect to the heavy-atom structure) between moduli of structure factors.

Equation (2.2.10.2) may be compared with Karle's (1983) qualitative rule: if the sign of

$$[(F_{\mathbf{h}})_{PH} - (F_{\mathbf{h}})_P][(F_{\mathbf{k}})_{PH} - (F_{\mathbf{k}})_P][(F_{\mathbf{l}})_{PH} - (F_{\mathbf{l}})_P]$$

is plus then the value of Φ_1 is estimated to be zero; if its sign is minus then the expected value of Φ_1 is close to π. In practice Karle's rule agrees with (2.2.10.2) only if the Cochran-type term in (2.2.10.2) may be neglected. Furthermore, (2.2.10.2) shows that large reliability values do not depend on the triple product of structure-factor differences, but on the triple product of pseudo-normalized differences. A series of papers (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Platas, 1995; Giacovazzo, Siliqi & Zanotti, 1995; Giacovazzo *et al.*, 1996) shows how equation (2.2.10.2) may be implemented in a direct procedure which proved to be able to estimate the protein phases correctly without any preliminary information on the heavy-atom substructure.

Combination of direct methods with the two-derivative case is also possible (Fortier *et al.*, 1984) and leads to more accurate estimates of triplet invariants provided experimental data are of sufficient accuracy.

2.2.10.4. *Integration of anomalous-dispersion techniques with direct methods*

If the frequency of the radiation is close to an absorption edge of an atom, then that atom will scatter the X-rays anomalously (see Chapter 2.4) according to $f = f' + if''$. This results in the breakdown of Friedel's law. It was soon realized that the Bijvoet difference could also be used in the determination of phases (Peerdeman & Bijvoet, 1956; Ramachandran & Raman, 1956; Okaya & Pepinsky, 1956). Since then, a great deal of work has been done both from algebraic (see Chapter 2.4) and from probabilistic points of view. In this section we are only interested in the second.

We will mention the following different cases:

(1) The OAS (one-wavelength anomalous scattering) case, also called SAS (single-wavelength anomalous scattering).

(2) The SIRAS (single isomorphous replacement combined with anomalous scattering) case. Typically, native protein and heavy-atom-derivative data are simultaneously available, with heavy atoms as anomalous scatterers.

(3) The MIRAS case, which generalizes the SIRAS case.

(4) The MAD case, a multiple-wavelength technique.

references