SAPI: Fan, H.-F. (1999). *Crystallographic software: teXsan for Windows.* http://www.rigaku.com/downloads/journal/Vol15.1.1998/texsan.pdf.

SnB: Weeks, C. M. & Miller, R. (1999). *The design and implementation of SnB version 2.0. J. Appl. Cryst.* **32**, 120–124.

SHELX97 and *SHELXS*: Sheldrick, G. M. (2000). *The SHELX home page*. http://shelx.uni-ac.gwdg.de/SHELX/.

SHELXD: 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.

SIR97: Altomare, A., Burla, M. C., Camalli, M., Cascarano, G. L., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G. & Spagna, R. (1999). *SIR97: a new tool for crystal structure determination and refinement. J. Appl. Cryst.* **32**, 115–119.

SIR2004: 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.

XTAL3.6.1: Hall, S. R., du Boulay, D. J. & Olthof-Hazekamp, R. (1999). *Xtal3.6 crystallographic software.* http://xtal.sourceforge.net/.

2.2.10. Direct methods in macromolecular crystallography

2.2.10.1. Introduction

The smallest protein molecules contain about 400 non-hydrogen atoms, so they cannot be solved *ab initio* by the algorithms specified in Sections 2.2.7 and 2.2.8. However, traditional direct methods are applied for:

(*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 ($d > 1.2$ Å). Because of series termination and other errors the electron-density map at $d > 1.2$ 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.

Quite relevant results have recently been obtained by integrating direct methods with some additional experimental information. In particular, we will describe the combination of direct methods with:

(*a*) direct-space techniques for the *ab initio* crystal structure solution of proteins;

(*b*) isomorphous-replacement (SIR–MIR) techniques;

(*c*) anomalous-dispersion (SAD–MAD) techniques;

(*d*) molecular replacement.

Point (*d*) will not be treated here, as it is described extensively in *IT* F, Part 13.

2.2.10.2. Ab initio crystal structure solution of proteins

Ab initio techniques do not require prior information of any atomic positions. The recent tremendous increase in computing speed led to direct methods evolving towards the rapid development of multisolution techniques. The new algorithms of the program *Shake-and-Bake* (Weeks *et al.*, 1994; Weeks & Miller, 1999; Hauptman *et al.*, 1999) allowed an impressive extension of the structural complexity amenable to direct phasing. In particular we mention: (*a*) the minimal principle (De Titta *et al.*, 1994), according to which the phase problem is considered as a constrained global optimization problem; (*b*) the refinement procedure, which alternately uses direct- and reciprocal-space techniques; and (*c*) the parameter-shift optimization technique (Bhuiya & Stanley, 1963), which aims at reducing the value of the minimal function (Hauptman, 1991; De Titta *et al.*, 1994). An effective variant of *Shake-and Bake* is *SHELXD* (Sheldrick, 1998) which cyclically alternates tangent refinement in reciprocal space with peak-list optimisation procedures in real space (Sheldrick & Gould, 1995). Detailed information on these programs is available in *IT* F (2001), Part 16.

A different approach is used by *ACORN* (Foadi *et al.*, 2000), which first locates a small fragment of the molecule (eventually by molecular-replacement techniques) to obtain a useful nonrandom starting set of phases, and then refines them by means of solvent-flattening techniques.

The program *SIR2004* (Burla *et al.*, 2005) uses the tangent formula as well as automatic Patterson techniques to obtain a first imperfect structural model; then direct-space techniques are used to refine the model. The Patterson approach is based on the use of the superposition minimum function (Buerger, 1959; Richardson & Jacobson, 1987; Sheldrick, 1992; Pavelcík, 1988; Pavelcík *et al.*, 1992; Burla *et al.*, 2004). It may be worth noting that even this approach is of multisolution type: up to 20 trial solutions are provided by using as pivots the highest maxima in the superposition minimum function.

It is today possible to solve structures up to 2500 non-hydrogen atoms in the asymmetric unit provided data at atomic (about 1 Å) resolution are available. Proteins with data at quasi-atomic resolution (say up to 1.5–1.6 Å) can also be solved, but with greater difficulties (Burla *et al.*, 2005). A simple evaluation of the potential of the *ab initio* techniques suggests that the structural complexity range and the resolution limits amenable to the *ab*

initio approach could be larger in the near future. The approach will profit by general technical advances like the increasing speed of computers and by the greater efficiency of informatic tools (*e.g.* faster Fourier-transform techniques). It could also profit from new specific crystallographic algorithms (for example, Oszlányi & Süto, 2004). It is of particular interest that extrapolating moduli and phases of nonmeasured reflections beyond the experimental resolution limit makes the *ab initio* phasing process more efficient, and leads to crystal structure solution even in cases in which the standard programs do not succeed (Caliandro *et al.*, 2005*a*,*b*). Moreover, the use of the extrapolated values improves the quality of the final electron-density maps and makes it easier to recognize the correct one among several trial structures.

2.2.10.3. Integration of direct methods with isomorphous replacement techniques

SIR–MIR cases are characterized by a situation in which there is one native protein and one or more heavy-atom substructures. In this situation the phasing procedure may be a two-step process: in the first stage the heavy-atom positions are identified by Patterson techniques (Rossmann, 1961; Okaya *et al.*, 1955) or by direct methods (Mukherjee *et al.*, 1989). In the second step the protein phases are estimated by exploiting the substructure information. Direct methods are able to contribute to both steps (see Sections 2.2.10.5 and 2.2.10.6). In Section 2.2.10.4 we show that direct methods are also able to suggest alternative one-step procedures by estimating structure invariants from isomorphous data.

2.2.10.4. SIR–MIR case: one-step procedures

The theoretical basis was established by Hauptman (1982*a*): his primary interest was to establish the two-phase and three-phase structure invariants by exploiting the experimental information provided by isomorphous data. The protein phases could be directly assigned *via* a tangent procedure.

Let us denote the modulus of the isomorphous difference as

$$\Delta F = |F_{\rm d}| - |F_{\rm p}|$$

where the subscripts d and p denote the derivative and the protein, respectively.

Denote also 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*)

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

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 **h**, **k**, **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:

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

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

$$Q_1 = 2[\sigma_3/\sigma_2^{3/2}]_{\rm p}|E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{l}}| + 2[\sigma_3/\sigma_2^{3/2}]_{\rm H}\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{l}}, \qquad (2.2.10.2)$$

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

$$\Delta = (F_{\rm d} - F_{\rm p})/(\textstyle\sum f_j^2)_{\rm H}^{1/2}.$$

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) algebraic rule: if the sign of $\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{l}}$ 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 similar mathematical approach has been applied to estimate quartet invariants *via* isomorphous data. The result may be summarized as follows: a quartet is a phase relationship of order $N_{\rm H}^{-1}$ (Giacovazzo & Siliqi, 1996*a*,*b*; see also Kyriakidis *et al.*, 1996), with reliability factor equal to

$$G = \frac{2\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{l}}\Delta_{\mathbf{h+k+l}}}{Q_4 N_{\mathrm{H}}} \times \left\{1 + \left(\Delta^2_{\mathbf{h+k}} - 1\right) + \left(\Delta^2_{\mathbf{h+l}} - 1\right) + \left(\Delta^2_{\mathbf{k+l}} - 1\right)\right\}, \quad (2.2.10.3)$$

where Q_4 is a suitable normalizing factor.

As previously stressed, equations (2.2.10.2) and (2.2.10.3) are valid if the lack of isomorphism and the errors in the measurements are assumed to be negligible. At first sight this approach seems more appealing than the traditional two-step procedures, however it did not prove to be competitive with them. The main reason is the absence in the Hauptman and Giacovazzo approaches of a probabilistic treatment of the errors: such a treatment, on the contrary, is basic for the traditional SIR–MIR techniques [see Blow & Crick (1959) and Terwilliger & Eisenberg (1987) for two related approaches].

The problem of the errors in the probabilistic scenario defined by the joint probability distribution functions approach has recently been overcome by Giacovazzo *et al.* (2001). In their probabilistic calculations the following assumptions were made:

$$|F_{\mathrm{d}j}| \exp(i\varphi_j) = |F_{\mathrm{p}}| \exp(i\varphi_{\mathrm{p}}) + F_{\mathrm{H}j} \exp(i\varphi_{\mathrm{H}j}) + |\mu_j| \exp(i\theta_j), \quad (2.2.10.4)$$

where j refers to the jth derivative. $|\mu_j| \exp(i\theta_j)$ is the error, which can include model as well as measurement errors.

A more realistic expression for the reliability factor G of triplet invariants is obtained by including the expression (2.2.10.4) in the probabilistic approach. Then the reliability parameter of the triplet invariants is transformed into (Giacovazzo *et al.*, 2001)

$$G = 2[\sigma_3/\sigma_2^{3/2}]_{\mathrm{p}} R_{\mathrm{p}1} R_{\mathrm{p}2} R_{\mathrm{p}3} + 2[\sigma_3/\sigma_2^{3/2}]_{\mathrm{H}} \frac{\Delta_1\Delta_2\Delta_3}{[1 + (\sigma^2_{\mu 1})_{\mathrm{H}}][1 + (\sigma^2_{\mu 2})_{\mathrm{H}}][1 + (\sigma^2_{\mu 3})_{\mathrm{H}}]}, \quad (2.2.10.5)$$

where $(\sigma^2_{\mu})_{\mathrm{H}} = |\mu|^2/(\Sigma f_j^2)_{\mathrm{H}}$.

Equation (2.2.10.5) suggests how the error influences the reliability of the triplet estimate: even quite a small value of $|\mu|^2$ may be critical if the scattering power of the heavy-atom substructure is a very small percentage of the derivative scattering power.

A one-step procedure has been implemented in a computer program (Giacovazzo *et al.*, 2002): it has been shown that the method is able to derive automatically, from the experimental data and without any user intervention, good quality (*i.e.* perfectly interpretable) electron-density maps.

2.2.10.5. SIR–MIR case: the two-step procedure. Finding the heavy-atom substructure by direct methods

The first trials for finding the heavy-atom substructure were based on the following assumption: the modulus of the isomorphous difference,

$$\Delta F = |F_{\mathrm{d}}| - |F_{\mathrm{p}}|,$$

is assumed at a first approximation as an estimate of the heavy-atom structure factor F_{H}. Perutz (1956) approximated $|F_{\mathrm{H}}|^2$ with the difference $(|F_{\mathrm{d}}|^2 - |F_{\mathrm{p}}|^2)$. Blow (1958) and Rossmann (1960) suggested a better approximation: $|F_{\mathrm{H}}|^2 \simeq |\Delta F|^2$. A deeper analysis was performed by Phillips (1966), Dodson & Vijayan (1971), Blessing & Smith (1999) and Grosse-Kunstleve & Brunger (1999). The use of direct methods requires the normalization of $|\Delta F|$ and application of the tangent formula (Wilson, 1978).

A sounder procedure has been suggested by Giacovazzo *et al.* (2004): they studied, for the SIR case, the joint probability distribution function

$$P(E_{\mathrm{H}}, E_{\mathrm{p}}, E_{\mathrm{d}})$$

under the following assumptions:

(*a*) the atomic positions of the native protein structure and the positions of the heavy atoms in the derivative structure are the primitive random variables of the probabilistic approach;

(*b*)

$$|F_{\mathrm{d}}| \exp(i\varphi_{\mathrm{d}}) = |F_{\mathrm{p}}| \exp(i\varphi_{\mathrm{p}}) + |F_{\mathrm{H}}| \exp(i\varphi_{\mathrm{H}}) + |\mu_{\mathrm{d}}| \exp(i\theta_{\mathrm{d}}) \quad (2.2.10.6)$$

is the structure factor of the derivative.

Then the conditional distribution $P(R_{\mathrm{H}}|R_{\mathrm{p}}, R_{\mathrm{d}})$ may be derived, from which $\langle R_{\mathrm{H}}|R_{\mathrm{p}}, R_{\mathrm{d}}\rangle$ may be obtained. In terms of structure factors

$$\langle |F_{\mathrm{H}}|^2 \rangle = \frac{\sum_{\mathrm{H}}}{\left(\sum_{\mathrm{H}} + \langle |\mu_{\mathrm{d}}|^2 \rangle\right)} \left[\langle |\mu_{\mathrm{d}}|^2 \rangle + \frac{\sum_{\mathrm{H}}}{\left(\sum_{\mathrm{H}} + \langle |\mu_{\mathrm{d}}|^2 \rangle\right)} \Delta^2_{\mathrm{iso}} \right]. \quad (2.2.10.7)$$

The effect of the errors on the evaluation of the moduli $|F_{\mathrm{H}}|^2$ may be easily derived: if $\langle |\mu_{\mathrm{d}}|^2 \rangle = 0$, equation (2.2.10.7) confirms Blow and Rossmann's approximation $\langle |F_{\mathrm{H}}|^2 \rangle \simeq |\Delta F|^2$. If $\langle |\mu_{\mathrm{d}}|^2 \rangle \neq 0$ Blow and Rossmann's estimate should be affected by a systematic error, increasing with $\langle |\mu_{\mathrm{d}}|^2 \rangle$.

2.2.10.6. SIR–MIR case: protein phasing by direct methods

Let us suppose that the various heavy-atom substructures have been determined. They may be used as additional prior information for a more accurate estimate of the φ_{p} values. To this purpose the distributions

$$P(E_{\mathrm{p}}, \mathbf{E}'_{\mathrm{d}}|\mathbf{E}'_{\mathrm{H}}) \equiv P(E_{\mathrm{p}}, E'_{\mathrm{d}1}, \ldots, E'_{\mathrm{d}n}|E'_{\mathrm{H}1}, \ldots, E'_{\mathrm{H}n}) \quad (2.2.10.8)$$

may be used under the assumption (2.2.10.6). $E'_{\mathrm{d}j}$ and $E'_{\mathrm{H}j}$, for $j = 1, \ldots, n$, are the structure factors of the jth derivative and of the jth heavy-atom substructure, respectively, both normalized with respect to the protein. Any joint probability density (2.2.10.8) may be reliably approximated by a multidimensional Gaussian distribution (Giacovazzo & Siliqi, 2002), from which the following conditional distribution is obtained:

$$P(\varphi_{\mathrm{p}}|R_{\mathrm{p}}, \mathbf{R}'_{\mathrm{d}}, \mathbf{E}'_{\mathrm{H}}) \simeq [2\pi I_0(G)]^{-1} \exp[\alpha_{\mathrm{p}} \cos(\varphi_{\mathrm{p}} - \theta_{\mathrm{p}})]$$

where θ_{p}, the expected value of φ_{p}, is given by

$$\tan\theta_{\mathrm{p}} = \frac{\sum_{j=1}^{n} G_j \sin\varphi_{\mathrm{H}j}}{\sum_{j=1}^{n} G_j \cos\varphi_{\mathrm{H}j}} = \frac{T}{B}$$

and $G_j = 2|F_{\mathrm{H}j}|\Delta F/\mu_j^2$.

$\alpha_{\mathrm{p}} = (T^2 + B^2)^{1/2}$ is the reliability factor of the phase estimate. A robust phasing procedure has been established which, starting from the observed moduli $|F_{\mathrm{p}}|, |F_{\mathrm{d}j}|, j = 1, \ldots, n$, is able to automatically provide, without any user intervention, a high-quality electron-density map of the protein (Giacovazzo *et al.*, 2002).

2.2.10.7. 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.

SAD (single anomalous dispersion) and MAD (multiple anomalous dispersion) techniques can be used. Both are characterized by one protein structure and one anomalous-scatterer substructure. The experimental diffraction data differ only because of the different anomalous scattering (not because of different anomalous-scatterer substructures). In the MAD case the anomalous-scatterer substructure is in some way 'overdetermined' by the data and, therefore, it is more convenient to use a two-step procedure: first define the positions of the anomalous scatterers, and then estimate the protein phase values. For completeness, we describe the one-step procedures in Section 2.2.10.8. These are based on the estimation of the structure invariants and on the application of the tangent formula. The two-step procedures are described in the Sections 2.2.10.9 and 2.2.10.10.

2.2.10.8. The SAD case: the one-step procedures

Probability distributions of diffraction intensities and of selected functions of diffraction intensities for dispersive structures have been given by various authors [Parthasarathy & Srinivasan (1964), see also Srinivasan & Parthasarathy (1976) and relevant literature cited therein]. We describe here some probabilistic formulae for estimating invariants of low order.

(*a*) *Estimation of two-phase structure invariants.* The conditional probability distribution of $\Phi = \varphi_{\mathbf{h}} + \varphi_{-\mathbf{h}}$ given $R_{\mathbf{h}}$ and $G_{\mathbf{h}}$ (normalized moduli of $F_{\mathbf{h}}$ and $F_{-\mathbf{h}}$, respectively) (Hauptman, 1982*b*; Giacovazzo, 1983*b*) is

$$P(\Phi|R_{\mathbf{h}}, G_{\mathbf{h}}) \simeq [2\pi I_0(Q)]^{-1} \exp[Q\cos(\Phi - q)], \qquad (2.2.10.9)$$

where

$$\begin{aligned} Q &= \frac{2R_{\mathbf{h}}G_{\mathbf{h}}}{\sqrt{c}}[c_1^2 + c_2^2]^{1/2}, \\ \cos q &= \frac{c_1}{[c_1^2 + c_2^2]^{1/2}}, \qquad \sin q = \frac{c_2}{[c_1^2 + c_2^2]^{1/2}}, \\ c_1 &= \sum_{j=1}^{N}(f_j'^2 - f_j''^2)/\sum, \\ c_2 &= 2\sum_{j=1}^{N} f_j' f_j''/\sum, \\ c &= [1 - (c_1^2 + c_2^2)]^2, \\ \sum &= \sum_{j=1}^{N}(f_j'^2 + f_j''^2). \end{aligned}$$

q is the most probable value of Φ: a large value of the parameter Q suggests that the phase relation $\Phi = q$ is reliable. Large values of Q are often available in practice: q, however, may be considered an estimate of $|\Phi|$ rather than of Φ because the enantiomorph is not fixed in (2.2.10.9). A formula for the estimation of Φ in centrosymmetric structures has been provided by Giacovazzo (1987).

(*b*) *Estimation of triplet invariants.* Kroon *et al.* (1977) first incorporated anomalous diffraction in order to estimate triplet invariants. Their work was based on an analysis of the complex double Patterson function. Subsequent probabilistic considerations (Heinermann *et al.*, 1978) confirmed their results, which can be so expressed:

$$\sin\bar{\Phi} = \frac{|\tau|^2 - |\bar{\tau}|^2}{4\tau''[\frac{1}{2}(|\tau|^2 + |\bar{\tau}|^2) - |\tau''|^2]^{1/2}}, \qquad (2.2.10.10)$$

where $(\mathbf{h} + \mathbf{k} + \mathbf{l} = 0)$,

$$\begin{aligned} \tau &= E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{l}} = R_{\mathbf{h}}R_{\mathbf{k}}R_{\mathbf{l}}\exp(i\Phi_{\mathbf{h},\mathbf{k}}), \\ \bar{\tau} &= E_{-\mathbf{h}}E_{-\mathbf{k}}E_{-\mathbf{l}} = G_{\mathbf{h}}G_{\mathbf{k}}G_{\mathbf{l}}\exp(i\Phi_{\bar{\mathbf{h}},\bar{\mathbf{k}}}), \\ \bar{\Phi} &= \tfrac{1}{2}(\Phi_{\mathbf{h},\mathbf{k}} - \Phi_{\bar{\mathbf{h}},\bar{\mathbf{k}}}), \end{aligned}$$

and τ'' is the contribution of the imaginary part of τ, which may be approximated in favourable conditions by

$$\begin{aligned} \tau'' = 2f''[&f_{\mathbf{h}}'f_{\mathbf{k}}' + f_{\mathbf{h}}'f_{\mathbf{l}}' + f_{\mathbf{k}}f_{\mathbf{l}}] \\ &\times [1 + S(R_{\mathbf{h}}^2 + R_{\mathbf{k}}^2 + R_{\mathbf{l}}^2 - 3)], \end{aligned}$$

where S is a suitable scale factor.

Equation (2.2.10.10) gives two possible values for $\bar{\Phi}$ (Φ and $\pi - \Phi$). Only if $R_{\mathbf{h}}R_{\mathbf{k}}R_{\overline{\mathbf{h}+\mathbf{k}}}$ is large enough may this phase ambiguity be resolved by choosing the angle nearest to zero.

The evaluation of triplet phases by means of anomalous dispersion has been further pursued by Hauptman (1982*b*) and independently by Giacovazzo (1983*b*). Owing to the breakdown of Friedel's law there are eight distinct triplet invariants which can contemporaneously be exploited:

$$\begin{aligned} \Phi_1 &= \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}}, & \Phi_2 &= -\varphi_{-\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}} \\ \Phi_3 &= \varphi_{\mathbf{h}} - \varphi_{-\mathbf{k}} + \varphi_{\mathbf{l}}, & \Phi_4 &= \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} - \varphi_{-\mathbf{l}} \\ \Phi_5 &= \varphi_{-\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{-\mathbf{l}}, & \Phi_6 &= -\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{-\mathbf{l}} \\ \Phi_7 &= \varphi_{-\mathbf{h}} - \varphi_{\mathbf{k}} + \varphi_{-\mathbf{l}}, & \Phi_8 &= \varphi_{-\mathbf{h}} + \varphi_{-\mathbf{k}} - \varphi_{\mathbf{l}}. \end{aligned}$$

Given

$$\begin{aligned} R_1 &= |E_{\mathbf{h}}| & R_2 &= |E_{\mathbf{k}}| & R_3 &= |E_{\mathbf{h}+\mathbf{k}}| \\ G_1 &= |E_{-\mathbf{h}}| & G_2 &= |E_{-\mathbf{k}}| & R_3 &= |E_{-\mathbf{h}-\mathbf{k}}| \\ \varphi_1 &= \varphi_{\mathbf{h}} & \varphi_2 &= \varphi_{\mathbf{k}} & \varphi_3 &= \varphi_{\mathbf{h}+\mathbf{k}} \\ \psi_1 &= \varphi_{-\mathbf{h}} & \psi_2 &= \varphi_{-\mathbf{k}} & \psi_3 &= \varphi_{-\mathbf{h}-\mathbf{k}}, \end{aligned}$$

Hauptman and Giacovazzo found the following conditional distribution:

$$P(\Phi|R_j, G_j, j = 1, 2, 3) \simeq \left[2\pi I_0(\Omega)\right]^{-1}\exp[\Omega\cos(\Phi - \omega)]. \qquad (2.2.10.11)$$

The definitions of Ω and ω are rather extensive and so the reader is referred to the published papers. We only add that Ω is always positive and that ω, the expected value of Φ, may lie anywhere between 0 and 2π. Understanding the role of the various parameters in equation (2.2.10.11) is not easy. Giacovazzo *et al.* (2003) found an equivalent simpler expression from which interpretable estimates of the parameters were obtained. In the same paper the limitations of the approach (*versus* the two-step procedures) were clarified.

2.2.10.9. SAD–MAD case: the two-step procedures. Finding the anomalous-scatterer substructure by direct methods

The anomalous-scatterer substructure is traditionally determined by the techniques suggested by Karle and Hendrickson (Karle, 1980*b*; Hendrickson, 1985; Pähler *et al.*, 1990; Terwilliger, 1994). The introduction of selenium into proteins as selenomethionine encouraged the second-generation direct methods programs [*Shake and Bake* by Miller *et al.* (1994); *Half bake* by Sheldrick (1998); *SIR2000-N* by Burla *et al.* (2001); *ACORN* by Foadi *et al.* (2000)] to locate Se atoms. Since the number of Se atoms may be quite large (up to 200), direct methods rather than Patterson techniques seem to be preferable. *Shake and Bake*, *Half Bake* and *ACORN* obtain the coordinates of the anomalous scatterers from a single-wavelength set of data. When more sets of diffraction data are available the solutions obtained by the other sets are used to confirm the correct solution.

A different approach has been suggested in two recent papers (Burla *et al.*, 2002; Burla, Carrozzini *et al.*, 2003): the estimates of the amplitudes of the structure factors of the anomalously scattering substructure are derived, *via* the rigorous method of the joint probability distribution functions, from the experimental diffraction moduli relative to n wavelengths. To do that, first the joint distribution

$$\begin{aligned} P_n = P(&A_{\rm oa}, A_1^+, A_2^+, \ldots, A_n^+, A_1^-, A_2^-, \ldots, A_n^-, \\ &B_{\rm oa}, B_1^+, B_2^+, \ldots, B_n^+, B_1^-, B_2^-, \ldots, B_n^-) \\ &= \pi^{-(2n+1)}(\det\, K)^{1/2} \exp(-\tfrac{1}{2}\mathbf{T}^T\mathbf{K}^{-1}\mathbf{T}) \end{aligned}$$

is calculated, where $A_{\rm oa}, B_{\rm oa}, E_{\rm oa}, A_i^+, B_i^+, A_i^-, B_i^-$ are the real and imaginary components of $E_{\rm oa}$, E_i^+, E_i^-, respectively, $\mathbf{K}$ is a symmetric square matrix of order $(4n + 2)$, $\mathbf{K}^{-1} = \{\lambda_{ij}\}$ is its inverse, and $\mathbf{T}$ is a suitable vector with components defined in terms of the variables $A_{\rm oa}, A_1^+, A_2^+, \ldots, B_n^-$. $E_{\rm oa}$ is the normalized structure factor of the anomalous scatterer substructure calculated by neglecting anomalous scattering components. Then the conditional distribution

$$P(R_{\rm oa}|R_1, \ldots, R_n, G_1, \ldots, G_n)$$

is derived, from which

$$\langle R_{\rm oa}|R_1, \ldots, G_n\rangle = \tfrac{1}{2}(\pi/\lambda_{11})^{1/2}[1 + 4X^2/(\pi\lambda_{11})]^{1/2} \quad (2.2.10.12)$$

is obtained, where

$$\begin{aligned} X^2 &= Q_1^2 + Q_2^2 \\ Q_1 &= \lambda_{12}R_1 + \lambda_{13}R_2 + \ldots + \lambda_{1,n+1}R_n + \lambda_{1,n+2}G_1 + \ldots \\ &\quad + \lambda_{1,2n+1}G_n \\ Q_2 &= \lambda_{1,2n+3}R_1 + \lambda_{1,2n+4}R_2 + \ldots + \lambda_{1,3n+2}R_n + \ldots - \lambda_{1,3n+3}G_1 \\ &\quad - \ldots - \lambda_{1,4n+2}G_n. \end{aligned}$$

The standard deviation of the estimate is also calculated:

$$\sigma_{R_{\rm oa}} = \left[\langle R_{\rm oa}^2|\ldots\rangle - \langle R_{\rm oa}|\ldots\rangle^2\right]^{1/2} = \left[\left(1 - \frac{\pi}{4}\right)\lambda_{11}^{-1}\right]^{1/2},$$

from which

$$\frac{\langle R_{\rm oa}|\ldots\rangle}{\sigma_{R_{\rm oa}}} = \left[\frac{(\pi/4) + (X^2)/\lambda_{11}}{1 - (\pi/4)}\right]^{1/2}. \quad (2.2.10.13)$$

The advantage of the above approach is that the estimates can simultaneously exploit both the anomalous and the dispersive differences. The computing procedure proposed by Burla, Carrozzini *et al.* (2003) is the following:

(i) The sets S_j, $j = 1, \ldots, n$, of the observed magnitudes (say $|F^+|$, $|F^-|$) are stored for all the n wavelengths.

(ii) The Wilson method is applied to put the sets S_j on their absolute scales.

(iii) Equations (2.2.10.12) and (2.2.10.13) are applied to obtain the values $\langle R_{\rm oa}|\ldots\rangle$ and $\langle R_{\rm oa}|\ldots\rangle/\sigma_{R_{\rm oa}}$.

(iv) The triplet invariants involving the reflections with the highest $\langle R_{\rm oa}|\ldots\rangle/\sigma_{R_{\rm oa}}$ values are evaluated and the tangent formula is applied *via* a random starting approach.

(v) The direct-space refinement techniques of *SIR2002* (Burla, Camalli *et al.*, 2003) are used to extend the phase information to a larger set of reflections: only 30% of the reflections with the smallest values of $\langle R_{\rm oa}|\ldots\rangle$ remain unphased. Automatic cycles of least-squares refinement improve the substructure model provided by the trial solutions.

(vi) Suitable figures of merit are used to recognize the correct substructure models.

The application of the above procedure to several MAD cases showed that the various wavelength combinations are not equally informative. A criterion based on the correlation among the various $\Delta_{\rm ano}$ values was also provided (see also Schneider & Sheldrick, 2002) for predicting the most informative combinations.

2.2.10.10. SAD–MAD case: protein phasing by direct methods

Once the anomalous-scatterer substructure has been found, the corresponding structure factors $E_{a1}^+, \ldots, E_{an}^+, E_{a1}^-, \ldots, E_{an}^-$ are known in modulus and phase. Then the conditional joint probability distribution

$$P\left(E_1^+, \ldots, E_n^+, E_1^-, \ldots, E_n^-|E_{a1}^+, \ldots, E_{an}^+, E_{a1}^-, \ldots, E_{an}^-\right)$$

may be calculated (Giacovazzo & Siliqi, 2004), from which the conditional distribution

$$P\left(\varphi_1^+|E_{ai}^+, E_{ai}^-, R_i, G_i, i = 1, \ldots, 2\right)$$

may be derived.

It has been shown that the most probable phase of φ_1^+, say θ_1^+, is the phase of the vector

$$\begin{aligned} &\sum_{j=1}^{n}\left[w_j^+E_{aj}^+ + w_j^-E_{aj}^{-*}\right] \\ &\quad + \sum_{j,p=1,p>j}^{n}\left[w_{jp}(E_{aj}^+ - E_{ap}^+) + w_{n+j,n+p}\left(E_{aj}^{-*} - E_{ap}^{-*}\right)\right] \\ &\quad + \sum_{j,p=1}^{n} w_{j,n+p}\left(E_{aj}^+ - E_{ap}^{-*}\right)] \end{aligned} \quad (2.2.10.14)$$

and the reliability parameter of the phase estimate is nothing other than the modulus of (2.2.10.14). The first term in (2.2.10.14) is a Sim-like contribution; the other terms, through the weights w, take into account the errors and the experimental differences $(R_j - R_p)$, $(G_j - G_p)$ and $(R_j - G_p)$.

References

Allegra, G. (1979). *Derivation of three-phase invariants from the Patterson function. Acta Cryst.* A**35**, 213–220.

Altomare, A., Burla, M. C., Camalli, M., Cascarano, G. L., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G. & Spagna, R. (1999). *SIR97: a new tool for crystal structure determination and refinement. J. Appl. Cryst.* **32**, 115–119.

Anzenhofer, K. & Hoppe, W. (1962). *Phys. Verh. Mosbach.* **13**, 119.

Ardito, G., Cascarano, G., Giacovazzo, C. & Luić, M. (1985). *1-Phase seminvariants and Harker sections. Z. Kristallogr.* **172**, 25–34.

Argos, P. & Rossmann, M. G. (1980). *Molecular replacement method.* In *Theory and Practice of Direct Methods in Crystallography,* edited by M. F. C. Ladd & R. A. Palmer, pp. 381–389. New York: Plenum.

Avrami, M. (1938). *Direct determination of crystal structure from X-ray data. Phys. Rev.* **54**, 300–303.

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.

Banerjee, K. (1933). *Determination of the signs of the Fourier terms in complete crystal structure analysis. Proc. R. Soc. London Ser. A*, **141**, 188–193.

Bertaut, E. F. (1955*a*). *La méthode statistique en cristallographie. I. Acta Cryst.* **8**, 537–543.

Bertaut, E. F. (1955*b*). *La méthode statistique en cristallographie. II. Quelques applications. Acta Cryst.* **8**, 544–548.

Bertaut, E. F. (1960). *Ordre logarithmique des densités de répartition. I. Acta Cryst.* **13**, 546–552.

Beurskens, P. T., Beurskens, G., de Gelder, R., Garcia-Granda, S., Gould, R. O., Israel, R. & Smits, J. M. M. (1999). *The DIRDIF-99 program system.* Crystallography Laboratory, University of Nijmegen, The Netherlands.

Beurskens, P. T., Gould, R. O., Bruins Slot, H. J. & Bosman, W. P. (1987). *Translation functions for the positioning of a well oriented molecular fragment. Z. Kristallogr.* **179**, 127–159.