$$\boldsymbol{A}\boldsymbol{\Phi} = \boldsymbol{C}, \tag{2.2.8.3}$$

giving the least-squares solution

$$\boldsymbol{\Phi} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{C}. \tag{2.2.8.4}$$

When approximate phases are available, the nearest integers may be found and equations (2.2.8.3) and (2.2.8.4) constitute the basis for further refinement.

Modified tangent procedures are also used, such as (Sint & Schenk, 1975; Busetta, 1976)

$$\tan\varphi_{\mathbf{h}} \simeq \frac{\sum_j G_{\mathbf{h},\mathbf{k}_j}\sin(\varphi_{\mathbf{k}_j}+\varphi_{\mathbf{h}-\mathbf{k}_j}-\Delta_j)}{\sum G_{\mathbf{h},\mathbf{k}_j}\cos(\varphi_{\mathbf{k}_j}+\varphi_{\mathbf{h}-\mathbf{k}_j}-\Delta_j)},$$

where Δ_j is an estimate for the triplet phase sum $(\varphi_{\mathbf{h}}-\varphi_{\mathbf{k}_j}-\varphi_{\mathbf{h}-\mathbf{k}_j})$.

(5) *Techniques based on the positivity of Karle–Hauptman determinants*

(The main formulae have been briefly described in Section 2.2.5.7.) The maximum determinant rule has been applied to solve small structures (de Rango, 1969; Vermin & de Graaff, 1978) *via* determinants of small order. It has, however, been found that their use (Taylor *et al.*, 1978) is not of sufficient power to justify the larger amount of computing time required by the technique as compared to that required by the tangent formula.

(6) *Tangent techniques using simultaneously triplets, quartets, . . .*

The availability of a large number of phase relationships, in particular during the first stages of a direct procedure, makes the phasing process easier. However, quartets are sums of two triplets with a common reflection. If the phase of this reflection (and/or of the other cross terms) is known then the quartet probability formulae described in Section 2.2.5.5 cannot hold. Similar considerations may be made for quintet relationships. Thus triplet, quartet and quintet formulae described in the preceding paragraphs, if used without modifications, will certainly introduce systematic errors in the tangent refinement process.

A method which takes into account correlation between triplets and quartets has been described (Giacovazzo, 1980*c*) [see also Freer & Gilmore (1980) for a first application], according to which

$$\tan\varphi_{\mathbf{h}} \simeq \frac{\sum_{\mathbf{k}} G\sin(\varphi_{\mathbf{k}}+\varphi_{\mathbf{h}-\mathbf{k}}) - \sum_{\mathbf{k},\mathbf{l}} G'\sin(\varphi_{\mathbf{k}}+\varphi_{\mathbf{l}}+\varphi_{\mathbf{h}-\mathbf{k}-\mathbf{l}})}{\sum_{\mathbf{k}} G\cos(\varphi_{\mathbf{k}}+\varphi_{\mathbf{h}-\mathbf{k}}) - \sum_{\mathbf{k},\mathbf{l}} G'\cos(\varphi_{\mathbf{k}}+\varphi_{\mathbf{l}}+\varphi_{\mathbf{h}-\mathbf{k}-\mathbf{l}})},$$

where G' takes into account both the magnitudes of the cross terms of the quartet and the fact that their phases may be known.

(7) *Integration of Patterson techniques and direct methods (Egert & Sheldrick, 1985) [see also Egert (1983, and references therein)]*

A fragment of known geometry is oriented in the unit cell by real-space Patterson rotation search (see Chapter 2.3) and its position is found by application of a translation function (see Section 2.2.5.4 and Chapter 2.3) or by maximizing the weighted sum of the cosines of a small number of strong translation-sensitive triple phase invariants, starting from random positions. Suitable FOMs rank the most reliable solutions.

(8) *Maximum entropy methods*

A common starting point for all direct methods is a stochastic process according to which crystal structures are thought of as being generated by randomly placing atoms in the asymmetric unit of the unit cell according to some *a priori* distribution. A non-uniform prior distribution of atoms $p(\mathbf{r})$ gives rise to a source of random atomic positions with entropy (Jaynes, 1957)

$$H(p) = -\int_V p(\mathbf{r})\log p(\mathbf{r})\,\mathrm{d}\mathbf{r}.$$

The maximum value $H_{\max} = \log V$ is reached for a uniform prior $p(\mathbf{r}) = 1/V$.

The strength of the restrictions introduced by $p(\mathbf{r})$ is not measured by $H(p)$ but by $H(p) - H_{\max}$, given by

$$H(p) - H_{\max} = -\int_V p(\mathbf{r})\log[\,p(\mathbf{r})/m(\mathbf{r})]\,\mathrm{d}\mathbf{r},$$

where $m(\mathbf{r}) = 1/V$. Accordingly, if a prior prejudice $m(\mathbf{r})$ exists, which maximizes H, the revised relative entropy is

$$S(p) = -\int_V p(\mathbf{r})\log[\,p(\mathbf{r})/m(\mathbf{r})]\,\mathrm{d}\mathbf{r}.$$

The maximization problem was solved by Jaynes (1957). If $G_j(p)$ are linear constraint functionals defined by given constraint functions $C_j(\mathbf{r})$ and constraint values c_j, *i.e.*

$$G_j(p) = \int_V p(\mathbf{r})C_j(\mathbf{r})\,\mathrm{d}\mathbf{r} = c_j,$$

the most unbiased probability density $p(\mathbf{r})$ under prior prejudice $m(\mathbf{r})$ is obtained by maximizing the entropy of $p(\mathbf{r})$ relative to $m(\mathbf{r})$. A standard variational technique suggests that the constrained maximization is equivalent to the unconstrained maximization of the functional

$$S(p) + \sum_j \lambda_j G_j(p),$$

where the λ_j's are Lagrange multipliers whose values can be determined from the constraints.

Such a technique has been applied to the problem of finding good electron-density maps in different ways by various authors (Wilkins *et al.*, 1983; Bricogne, 1984; Navaza, 1985; Navaza *et al.*, 1983).

Maximum entropy methods are strictly connected with traditional direct methods: in particular it has been shown that:

(*a*) the maximum determinant rule (see Section 2.2.5.7) is strictly connected (Britten & Collins, 1982; Piro, 1983; Narayan & Nityananda, 1982; Bricogne, 1984);

(*b*) the construction of conditional probability distributions of structure factors amounts precisely to a reciprocal-space evaluation of the entropy functional $S(p)$ (Bricogne, 1984).

Maximum entropy methods are under strong development: important contributions can be expected in the near future even if a multipurpose robust program has not yet been written.

2.2.9. Some references to direct-methods packages: the small-molecule case

Some references for direct-methods packages are given below. Other useful packages using symbolic addition or multisolution procedures do exist but are not well documented.

CRUNCH: Gelder, R. de, de Graaff, R. A. G. & Schenk, H. (1993). *Automatic determination of crystal structures using Karle–Hauptman matrices. Acta Cryst.* A**49**, 287–293.

DIRDIF: 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.

MITHRIL: Gilmore, C. J. (1984). *MITHRIL. An integrated direct-methods computer program. J. Appl. Cryst.* **17**, 42–46.

MULTAN88: Main, P., Fiske, S. J., Germain, G., Hull, S. E., Declercq, J.-P., Lessinger, L. & Woolfson, M. M. (1999). *Crystallographic software: teXsan for Windows.* http://www.rigaku.com/downloads/journal/Vol15.1.1998/texsan.pdf.

PATSEE: Egert, E. & Sheldrick, G. M. (1985). *Search for a fragment of known geometry by integrated Patterson and direct methods. Acta Cryst.* A**41**, 262–268.

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*