Acta Cryst. (1983). A**39**, 61–68

Information Theory and the 'Phase Problem' in Crystallography

BY OSCAR E. PIRO*†

Department of Biophysics and Theoretical Biology, The University of Chicago, Illinois 60637, *USA*

(*Received* 21 *December* 1981; *accepted* 29 *July* 1982)

Abstract

The 'phase problem' is cast in terms of information theory. Constrained maximum-entropy inference is used to obtain the statistical structure of the largest possible set of electron density functions compatible with conditions that underlie the derivation of the Sayre–Hughes equation. As a consequence, the information provided by the knowledge of some structure factors is quantitatively expressed. The most unbiased prediction for the phases of structure factors whose moduli are known leads to the statement of the 'minimum added information rule' which corresponds with Tsoucaris's 'maximum determinant rule' [Tsoucaris (1970). *Acta Cryst.* A**26**, 492–499]. Expansion of the information in multiplets followed by the application of the proposed rule leads to the current formulae of direct methods. The information content of these formulae is discussed, and its dependence upon the magnitude of the structure factors and *a priori* structural knowledge is emphasized.

1. Introduction

The purpose of this paper is to present a derivation of the current formulation of direct methods using information theory as a conceptual framework. As a consequence of this approach, we will give a measure of the information in the phase determination procedure.

Information theory, initiated mainly by the work of C. E. Shannon (Shannon & Weaver, 1949), is based on statistical mathematics. Though it was initially developed to be used in problems of communication, its concepts and methods have been applied to other areas of science (Brillouin, 1962). Among the previous applications of information theory to crystallography we should mention the work by Diamond (1963) who introduced a measure (in 'bits') of the information contained in inequality of Karle–Hauptman determinants. Hosoya & Tokonami (1967) considered the estimation of the conformational entropy of an essentially one-dimensional real structure and the removal of structural uncertainty during crystal structure determination by the information contained in the reflection intensities and in the Patterson peaks. de Rango, Tsoucaris & Zelwer (1974) stressed the relation between the efficiency of the probability laws for phase determination and information theory.

Shannon's work established clearly the connection between information and entropy. This connection leads to the guiding principle that the probability distribution of a set of physical magnitudes related to a system that is not amenable to a complete experimental determination and that has maximum entropy subject to whatever is known provides the most unbiased representation of the system.

The principle referred to above has been used previously as a statistical method for prediction of phases of structure factors in crystallography (Piro & Podjarny, 1978, 1979). By applying the constrained maximum-entropy procedure they obtained a quantitative expression for the information added by a set of structure factors. When the moduli of these structure factors are experimentally known, the most unbiased estimation of the phases leads to the 'minimum added information rule' with respect to which Tsoucaris's 'maximum determinant rule' is a particular case (Piro, 1977; Piro & Podjarny, 1978, 1979). The expansion of the information in multiplets, in conjunction with the application of the minimum added information rule, allows several existing phase-determination procedures to be obtained from the viewpoint of information theory.

In the present work we derive the information content of some invariants and quantitate the gain in information due to both *a priori* structural knowledge and the experimental knowledge of the signs of some triplet products.

2. Multivariate joint probability distribution of normalized structure factors

We want to investigate the statistical structure of the continuous electron-density functions $\rho_E(\mathbf{r})$ having the continuous argument $\mathbf{r}$, using the simpler multivariate

* Supported by a fellowship from the CONICET, Argentina.

† Present address: Departamento de Fisica, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Calle 115 esquina 49, CC 67, La Plata 1900, Argentina.

0567-7394/83/010061-08$01.50

joint probability distribution of a discrete number m of continuous structure factors:

$$P(E_{\mathbf{h}_1}, E_{\mathbf{h}_2}, \ldots, E_{\mathbf{h}_m}) = P(\mathbf{E}),$$

where $\mathbf{E}$ is a column vector whose components are the m structure factors.

For simplicity we shall assume that all atoms are identical and therefore the elements of $\mathbf{E}$ are given by

$$E_{\mathbf{h}} = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \exp(2\pi i \mathbf{h}.\mathbf{r}_j); \tag{2.1}$$

$\mathbf{h}$: reciprocal-lattice vector, $\mathbf{r}_j$: vector of coordinates of the jth atom, N: number of atoms in the unit cell.

The unitary structure factors are given by

$$U_{\mathbf{h}} = E_{\mathbf{h}}/\sqrt{N}. \tag{2.2}$$

$P(\mathbf{E})$ satisfies the normalization condition

$$\int_{-\infty}^{\infty} \ldots \int P(\mathbf{E}_{\mathbf{h}_1}, E_{\mathbf{h}_2}, \ldots, E_{\mathbf{h}_m})\, \mathrm{d}E_{\mathbf{h}_1}\, \mathrm{d}E_{\mathbf{h}_2} \ldots \mathrm{d}E_{\mathbf{h}_m} = \int P(\mathbf{E})\, \mathrm{d}^m \mathbf{E} = 1\dagger. \tag{2.3}$$

The entropy or lack of information associated with such distribution is (Shannon & Weaver, 1949):

$$H = -\int P(\mathbf{E}) \ln P(\mathbf{E})\, \mathrm{d}^m \mathbf{E} = -\langle \ln P \rangle. \tag{2.4}$$

The distribution function $P(\mathbf{E})$ corresponding to the maximum lack of information and compatible with such *a priori* structural knowledge as positivity and atomicity of the electron-density function, expressed in reciprocal space by the Sayre–Hughes equation (Sayre, 1952; Hughes, 1953):

$$\langle E_{\mathbf{h}_j} E^*_{\mathbf{h}_i} \rangle = U_{\mathbf{h}_j - \mathbf{h}_i}, \tag{2.5}$$

can be found by a mathematical technique usual in dealing with analogous problems that arise in communication, *i.e.* using Lagrange multipliers (Shannon & Weaver, 1949):

$$\delta\left\{H(P) + \alpha\left[\int P(\mathbf{E})\, \mathrm{d}^m \mathbf{E} - 1\right] - \sum_{i,j=1}^{m} \beta_{ij}\left[\int E_{\mathbf{h}_j} E^*_{\mathbf{h}_i} P(\mathbf{E})\, \mathrm{d}^m \mathbf{E} - U_{\mathbf{h}_j - \mathbf{h}_i}\right]\right\} = \int \left\{-\ln P(\mathbf{E}) - 1 + \alpha - \sum_{i,j=1}^{m} \beta_{ij} E^*_{\mathbf{h}_i} E_{\mathbf{h}_j}\right\} \delta P\, \mathrm{d}^m \mathbf{E} = 0. \tag{2.6}$$

† In $P1$ the notation $P(E_{\mathbf{h}_1}, \ldots, E_{\mathbf{h}_m})$, where the E's are complex variables ($E_{\mathbf{h}} = A_{\mathbf{h}} + iB_{\mathbf{h}}$), in fact represents the joint probability function $P(A_{\mathbf{h}_1}, B_{\mathbf{h}_1}, \ldots, A_{\mathbf{h}_m}, B_{\mathbf{h}_m})$ for the real and imaginary parts of the m normalized structure factors. Consequently $\mathrm{d}E_{\mathbf{h}} = \mathrm{d}A_{\mathbf{h}}\, \mathrm{d}B_{\mathbf{h}}$ for the differential elements in the integrations.

Since (2.6) must hold for every variation δP of the conditional distribution function, the latter should be of the form:

$$P(\mathbf{E}) = \exp(\alpha - 1) \exp\left(-\sum_{i,j=1}^{m} \beta_{ij} E^*_{\mathbf{h}_i} E_{\mathbf{h}_j}\right). \tag{2.7}$$

From the constraint (2.5), we can deduce the value of the $\boldsymbol{\beta}$ matrix $\{(\boldsymbol{\beta})_{ij} \equiv \beta_{ij}\}$:

$$\boldsymbol{\beta} = \begin{cases} \frac{1}{2}\mathbf{U}^{-1} & \text{centric case } (P\bar{1}) \\ \mathbf{U}^{-1} & \text{acentric case } (P1), \end{cases} \tag{2.8}$$

where $(\mathbf{U})_{ij} = U_{\mathbf{h}_i - \mathbf{h}_j}$ is a Karle–Hauptman matrix (see Appendix).

From the normalization condition (2.3), we find for $P(\mathbf{E})$ in the two cases:

$$P(\mathbf{E}) = \begin{cases} \dfrac{[\det(\mathbf{U}^{-1})]^{1/2}}{(2\pi)^{m/2}} \exp(-\frac{1}{2}\mathbf{E}^T \mathbf{U}^{-1} \mathbf{E}) & (P\bar{1} \text{ case}) \\ \dfrac{\det(\mathbf{U}^{-1})}{\pi^m} \exp(-\mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E}) & (P1 \text{ case}), \end{cases} \tag{2.9}$$

where the row vectors $\mathbf{E}^T$ and $\mathbf{E}^+$ are, respectively, the transposed and conjugated transposed of the column vector $\mathbf{E}$.

The expression (2.9) for the conditional joint probability distribution of m normalized structure factors was deduced by Tsoucaris (1970) using statistical arguments based on the *central limit theorem*.

From the standpoint of information theory, we can express the above result saying that, from all possible distributions $P(\mathbf{E})$ compatible with the crystallographic constraints (2.5), the Laplace–Gauss distribution corresponds to the largest set of different possible electron-density functions. Therefore, the knowledge of the correct function removes the largest possible lack of information.

It is worth while to note that the mathematical procedure followed above to obtain the distribution function $P(\mathbf{E})$ given by (2.9) is a general one. If other constraints in addition to (2.5) can be expressed in reciprocal space, then additional Lagrange multipliers can be used to find the new distribution function which produces the constrained maximum entropy.

3. Information. Rule of the minimum information

Let us take the $P1$ case in (2.9):

$$P(E) = C \exp(-\mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E}), \tag{3.1}$$

where C is a constant independent of the structure factors $(E_{\mathbf{h}_1}, E_{\mathbf{h}_2}, \ldots, E_{\mathbf{h}_m})$. If we lack information about the m structure factors, then the average uncertainty

removed when we consider the whole set of $\rho_E(\mathbf{r})$ functions whose statistical structure is given by (2.9) is

$$H_b = -\langle \ln P \rangle = -\ln C + \langle \mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E} \rangle$$
$$= -\ln C + m, \tag{3.2}$$

while the actual value for the entropy removed when we have complete knowledge about $\mathbf{E}$ (in modulus and phase) is

$$H_a = -\ln P = -\ln C + \mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E}. \tag{3.3}$$

Therefore, it is reasonable to associate the information gained by such knowledge to the positive definite quantity:

$$\text{Information} = \mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E}. \tag{3.4a}$$

Analogously, we obtain for the centrosymmetrical case:

$$\text{Information} = \tfrac{1}{2}\mathbf{E}^T \mathbf{U}^{-1} \mathbf{E}. \tag{3.4b}$$

The origin of the factor $\frac{1}{2}$ in (3.4*b*) can be traced to the fact that by considering the centrosymmetric crystal as a special case belonging to the $P1$ space group, the corresponding information content carried by the E's and given by (3.4*a*) is twice as much as the true information because only half of the features exhibited by the associated E map are independent in $P\bar{1}$.

According to (3.4) the most unbiased prediction for the phases (signs) of the m structure factors whose moduli are known leads to the following rule: 'the most probable phases (signs) are those that minimize the information'.

It has been shown that (Tsoucaris, 1970):

$$\mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E} = N\left[\frac{D_m - \Delta_{m+1}}{D_m}\right], \tag{3.5}$$

where $D_m[= \det(\mathbf{U})]$ and Δ_{m+1} are Karle–Hauptman determinants. Δ_{m+1} is $(1/N)$ times the determinant obtained from D_m adding as a last row the m-dimensional row vector $\mathbf{E}^+$ and as a last column the corresponding column vector $\mathbf{E}$, while the $(m+1, m+1)$ determinant element is set equal to N.

From (3.4) and (3.5) it can be seen that our rule of minimum added information leads to the Tsoucaris 'maximum determinant rule' (Tsoucaris, 1970; Piro, 1977; Piro & Podjarny, 1978, 1979).

4. Expansion of the information in multiplets. Information content of some invariants

Let us separate the Karle–Hauptman matrix $\mathbf{U}$ into two terms:

$$\mathbf{U} = \mathbf{I} + \mathbf{U}', \tag{4.1}$$

where $\mathbf{I}$ is the unit matrix. Expanding $\mathbf{U}^{-1}$ in powers of $\mathbf{U}'$:

$$\mathbf{U}^{-1} = \mathbf{I} - \mathbf{U}' + \mathbf{U}'^2 - \mathbf{U}'^3 + \ldots, \tag{4.2}$$

the equations (3.4) for the information give rise to a multiplet expansion:

$$\text{Information} = \mathbf{E}^+ \mathbf{I}\, \mathbf{E} - \mathbf{E}^+ \mathbf{U}' \mathbf{E} + \mathbf{E}^+ \mathbf{U}'^2 \mathbf{E} - \mathbf{E}^+ \mathbf{U}'^3 \mathbf{E} + \ldots$$
$$= \sum_{\mathbf{h}_i} |E_{\mathbf{h}_i}|^2 - \sum_{\mathbf{h}_i \neq \mathbf{h}_j} \sum E^*_{\mathbf{h}_i} U_{\mathbf{h}_i - \mathbf{h}_j} E_{\mathbf{h}_j}$$
$$+ \sum_{\substack{\mathbf{h}_i \neq \mathbf{h}_j \neq \mathbf{h}_k \\ \mathbf{h}_i \neq \mathbf{h}_k}} \sum \sum E^*_{\mathbf{h}_i} U_{\mathbf{h}_i - \mathbf{h}_j} U_{\mathbf{h}_j - \mathbf{h}_k} E_{\mathbf{h}_k} - \ldots. \tag{4.3}$$

Some conclusions can be drawn from (4.3):

(i) We can recognize the leading term as a 'Wilson-type' term. This term, in general, will be the largest in the expansion, thus emphasizing the greater information content of the large E's as compared with the information provided by the small E's. On the other hand, this term will be the only one in the expansion if we set to zero all correlations between different structure factors in (2.5). The resultant multivariate joint distribution corresponds to contributing 'white noise' to the electron density due to the m new added structure factors.

(ii) The next term is the first one in the expansion containing phase information. The rule of minimum added information leads to the fact that the most probable phase of a single large triplet product should be zero (modulo 2π). If it is different from zero, some peculiarity in the structure produces this unexpected behavior and the corresponding information will increase. We will discuss the quantification of this fact in the next section and a numerical example taken from a real case will be given in § 7.

On the other hand, if we apply the rule to the sum of all the triplet products, then the tangent formula (Karle & Karle, 1966) is obtained. To prove this, let us express the information as given by its first two terms in real form:

$$I = \sum_{\mathbf{h}} |E_{\mathbf{h}}|^2 - \sum_{\mathbf{h} \neq \mathbf{k}} \sum |E_{\mathbf{h}} U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \cos(\varphi_{\mathbf{h}-\mathbf{k}} + \varphi_{\mathbf{k}} - \varphi_{\mathbf{h}}). \tag{4.4}$$

Let us segregate the contribution to the information (4.4) containing a given phase $\varphi_{\mathbf{h}}$:

$$I = \sum_{\mathbf{h}} |E_{\mathbf{h}}|^2 - 2|E_{\mathbf{h}}| \times \cos\varphi_{\mathbf{h}} \sum_{\mathbf{k} \neq \mathbf{h}} |U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \cos(\varphi_{\mathbf{h}-\mathbf{k}} + \varphi_{\mathbf{k}})$$
$$- 2|E_{\mathbf{h}}| \sin\varphi_{\mathbf{h}} \sum_{\mathbf{k} \neq \mathbf{h}} |U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \sin(\varphi_{\mathbf{h}-\mathbf{k}} + \varphi_{\mathbf{k}})$$
$$- \sum_{\substack{\mathbf{h}' \neq \mathbf{h} \neq \mathbf{k} \\ \mathbf{h}' \neq \mathbf{k}}} \sum |E_{\mathbf{h}'} U_{\mathbf{h}'-\mathbf{k}} E_{\mathbf{k}}| \cos(\varphi_{\mathbf{h}'-\mathbf{k}} + \varphi_{\mathbf{k}} - \varphi_{\mathbf{h}'}). \tag{4.5}$$

The condition of minimum added information is obtained when $\partial I/\partial\varphi_{\mathbf{h}} = 0$, which applied to (4.5) leads to:

$$\tan\varphi_{\mathbf{h}} = \frac{\sum_{\mathbf{k}\neq\mathbf{h}} |U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \sin(\varphi_{\mathbf{h}-\mathbf{k}}+\varphi_{\mathbf{k}})}{\sum_{\mathbf{k}\neq\mathbf{k}} |U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \cos(\varphi_{\mathbf{h}-\mathbf{k}}+\varphi_{\mathbf{k}})}. \quad (4.6)$$

(iii) Let us consider the information approximately expressed by the first three terms of (4.3). The rule of minimum added information will tend to predict the phase of a large quartet $E^*_{\mathbf{h}_i} U_{\mathbf{h}_i-\mathbf{h}_j} U_{\mathbf{h}_j-\mathbf{h}_k} E_{\mathbf{h}_k}$ to be (*a*) in the neighborhood of zero if the modulus of one or more of the 'cross terms' $|E_{\mathbf{h}_j}|$, $|E_{\mathbf{h}_i-\mathbf{h}_j+\mathbf{h}_k}|$, $|E_{\mathbf{h}_i-\mathbf{h}_k}|$ is large, or (*b*) in the neighborhood of π if all of these cross terms are small. To prove this, let us consider a large quartet as proportional to the product of two triplet products, *e.g.*

$$(E^*_{\mathbf{h}_i} U_{\mathbf{h}_i-\mathbf{h}_j} E_{\mathbf{h}_j})(E^*_{\mathbf{h}_j} U_{\mathbf{h}_j-\mathbf{h}_k} E_{\mathbf{h}_k}).$$

If the phase of the large quartet is assumed to be π, it is likely that larger negative terms will contribute to the second summation of (4.3), especially so for large cross terms (like $E_{\mathbf{h}_j}$). Thus for large cross terms we expect the quartet phase to be in the neighborhood of zero. However, if the cross terms are small no such effect occurs and therefore the third summation in (4.3) predicts the phase of the quartet to be π. These conclusions are in close agreement with the findings of Schenk (1973, 1974) and Hauptman (1974) relative to the distribution and use of phase relationships among quartets of reflections.

A derivation similar to that given in (ii), but now considering the information in third-order approximation, leads to a 'generalized tangent formula':

$$\tan\varphi_{\mathbf{h}} = \left[\sum_{\mathbf{k}\neq\mathbf{h}} |U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \sin(\varphi_{\mathbf{h}-\mathbf{k}}+\varphi_{\mathbf{k}}) - \sum_{\mathbf{k}\neq\mathbf{h}\neq\mathbf{l}}^{\mathbf{k}\neq\mathbf{l}} |U_{\mathbf{h}-\mathbf{k}}U_{\mathbf{k}-\mathbf{l}}E_{\mathbf{l}}| \times \sin(\varphi_{\mathbf{h}-\mathbf{k}}+\varphi_{\mathbf{k}-\mathbf{l}}+\varphi_{\mathbf{l}})\right] \times \left[\sum_{\mathbf{k}\neq\mathbf{h}} |U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \cos(\varphi_{\mathbf{h}-\mathbf{k}}+\varphi_{\mathbf{k}}) - \sum_{\substack{\mathbf{k}\neq\mathbf{h}\neq\mathbf{l}\\ \mathbf{k}\neq\mathbf{l}}} |U_{\mathbf{h}-\mathbf{k}}U_{\mathbf{k}-\mathbf{l}}E_{\mathbf{l}}| \times \cos(\varphi_{\mathbf{h}-\mathbf{k}}+\varphi_{\mathbf{k}-\mathbf{l}}+\varphi_{\mathbf{l}})\right]^{-1}, \quad (4.7)$$

which incorporates in a consistent way phase relationships simultaneously involving triplet and quartet products.

5. Expected information to the second order of approximation (acentric case)

We saw in § 3 that when we lack any knowledge about the m structure factors [except for their correlations (2.5) which form the assumed known covariance matrix **U**], the expected information is:

$$\langle I\rangle = \langle \mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E}\rangle = m. \quad (5.1)$$

It is interesting to calculate, in a second-order approximation, the expected information when we know experimentally the moduli of the structure factors.

In order to average (4.4), the following relation must be taken into account:

$$\langle\cos(\varphi_{\mathbf{h}-\mathbf{k}}+\varphi_{\mathbf{k}}-\varphi_{\mathbf{h}})\rangle = \frac{I_1(K_{\mathbf{hk}})}{I_0(K_{\mathbf{hk}})}, \quad (5.2)$$

where I_0, I_1 are the zero- and first-order modified Bessel functions:

$$K_{\mathbf{hk}} = 2\sigma_3\,\sigma_2^{-3/2}|E_{\mathbf{h}}E_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}|,$$

$$\sigma_n = \sum_{j=1}^{N} f_j^n,$$

and f_j is the form factor of the jth atom (Germain, Main & Woolfson, 1970). Introducing (5.2) into (4.4), we obtain

$$\langle I\rangle = \sum_{\mathbf{h}} |E_{\mathbf{h}}|^2 - \sum_{\mathbf{h}\neq\mathbf{k}}\sum |E_{\mathbf{h}}U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \frac{I_1(K_{\mathbf{hk}})}{I_0(K_{\mathbf{hk}})}. \quad (5.3)$$

The average (5.2) could be deduced using the fact that the phase distribution of a triplet product $X = E_{-\mathbf{h}}E_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}$ (assuming known moduli for the factors) is a von Mises distribution (von Mises, 1918):

$$P(\varphi_{\mathbf{hk}}) = \frac{\exp[K'_{\mathbf{hk}}\cos(\varphi_{\mathbf{hk}}-q_{\mathbf{hk}})]}{2\pi I_0(K'_{\mathbf{hk}})}, \quad (5.4)$$

where

$$K'_{\mathbf{hk}}\exp(iq_{\mathbf{hk}}) = 2Q_{\mathbf{hk}}|E_{\mathbf{h}}E_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}|\exp(iq_{\mathbf{hk}}) = \frac{2\langle X\rangle|X|}{\sigma^2(X)}. \quad (5.5)$$

In the case of atoms randomly positioned, $2Q_{\mathbf{hk}}|E_{\mathbf{h}}E_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}|\exp(iq_{\mathbf{hk}})$ reduces to the real quantity:

$$2\sigma_3\,\sigma_2^{-3/2}|E_{\mathbf{h}}E_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \quad (q_{\mathbf{hk}}=0) \quad (5.6)$$

(Hendrickson & Lattman, 1970; Koenig, 1972, 1976; Heinerman, Krabbendam & Kroon, 1977).

Recently *a priori* structural information has been used in the phase probability distribution of triplet products and in the modification of the tangent formula (Main, 1976; Heinerman, 1977; Heinerman, Krabbendam & Kroon, 1977). The corresponding expected information under the condition of previous knowledge

of the diffraction intensities and of structural features is calculated as has been done before, but now, in general, $q_{\mathbf{hk}} \neq 0$:

$$\langle \cos \varphi_{\mathbf{hk}} \rangle = \frac{1}{2\pi I_o(K'_{\mathbf{hk}})} \int_{-\pi}^{\pi} \exp\left[K'_{\mathbf{hk}} \cos\left(\varphi_{\mathbf{hk}} - q_{\mathbf{hk}}\right)\right] \times \cos \varphi_{\mathbf{hk}} \, \mathrm{d}\varphi_{\mathbf{hk}} = \frac{I_1(K'_{\mathbf{hk}})}{I_0(K'_{\mathbf{hk}})} \cos q_{\mathbf{hk}}, \tag{5.7}$$

and

$$\langle I \rangle = \sum_{\mathbf{h}} |E_{\mathbf{h}}|^2 - \sum_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{h}} U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \frac{I_1(K'_{\mathbf{hk}})}{I_0(K'_{\mathbf{hk}})} \cos q_{\mathbf{hk}}. \tag{5.8}$$

From (5.3) and (5.8), we obtain the following expression for the gain in information due to *a priori* structural knowledge:

$$\text{Gain} = \sum_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{h}} U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \times \left\{ \frac{I_1(K_{\mathbf{hk}})}{I_0(K_{\mathbf{hk}})} - \frac{I_1(K'_{\mathbf{hk}})}{I_0(K'_{\mathbf{hk}})} \cos q_{\mathbf{hk}} \right\}. \tag{5.9}$$

6. Expected information to the second order of approximation (centric case)

In the centric case, the information is given by

$$I = \tfrac{1}{2} \mathbf{E}^T \mathbf{U}^{-1} \mathbf{E}, \tag{6.1}$$

and to the second order of approximation by

$$I = \tfrac{1}{2} \left\{ \sum_{\mathbf{h}} E_{\mathbf{h}}^2 - \sum_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{h}} U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \, s(\mathbf{hk}) \right\}, \tag{6.2}$$

where $s(\mathbf{hk})$ is the sign of the triplet product $E_{\mathbf{h}} U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}$.

Averaging (6.2) and remembering that

$$\langle s(\mathbf{hk}) \rangle = P(+) - P(-), \tag{6.3}$$

where, for randomly positioned atoms (Cochran & Woolfson, 1955),

$$P(\pm) = \tfrac{1}{2} \pm \tfrac{1}{2} \tanh\left(N^{-1/2} |E_{\mathbf{h}} E_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}|\right), \tag{6.4}$$

we obtain

$$I = \tfrac{1}{2} \left\{ \sum_{\mathbf{h}} E_{\mathbf{h}}^2 - \sum_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{h}} U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \times \tanh\left(N^{-1/2} |E_{\mathbf{h}} E_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}|\right) \right\}. \tag{6.5}$$

Let us consider the contribution to (6.2) and (6.5) due to an individual triplet product. If it is large and at the same time positive, as expected, its contribution to the information takes its minimum value, and therefore it does not tell us anything new. If it is large and negative, we can see from (6.2) that it adds an information larger than expected; there must be some structural peculiarity that produces such unexpected behavior.*

Recently, experimental measurement of the signs of large triplet products, based on the analysis of the distribution of the diffraction intensities about the three-beam point in a three-beam simultaneous diffraction experiment, have been successful on highly perfect germanium and relatively imperfect aluminum oxide crystals (Post, 1979). Although it is not yet possible to predict the utility of the technique applied to the 'mosaic' crystals usually found in crystal structure analysis, we can see, according to our previous discussion, the importance that the experimental knowledge of the negativity of some triplet products has in direct-methods calculations. We can easily measure the gain of information due to the experimental knowledge of negative triplets. From (6.2) and (6.5) we obtain:

$$\text{Gain}^{\text{exp}} = \tfrac{1}{2} \sum_{\mathbf{h}}' \sum_{\mathbf{k}}' |E_{\mathbf{h}} U_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}| \times \{1 + \tanh\left(N^{-1/2} |E_{\mathbf{h}} E_{\mathbf{h}-\mathbf{k}} E_{\mathbf{k}}|\right)\}, \tag{6.6}$$

where the sum is performed on the experimentally detected negative triplet products.

7. A simple case

In order to check quantitatively some aspects of the information-theory approach to the 'phase problem' developed in the previous sections, calculation of the information content of single triplet products in a specific case were performed. The test example taken was the centrosymmetric (010) projection of the crystal structure of 1,4-cyclohexanedione ($C_6H_8O_2$) at 133 K (Mossel & Romers, 1964). In this structure (space group $P2_1$) the molecules are stacked along **b** with their mean plane approximately parallel to the (010) plane as revealed by the Patterson projection along [010] calculated by the authors and which provided the knowledge of the rough molecular orientation. The particular orientation of the molecules in the crystal makes this compound interesting to analyze because its deviation from a random atom distribution should reflect itself markedly on the *a priori* expected information of single triplet products.

* This is a case similar to the information content of the outcomes during the throw of an uneven coin: p(head) $\gg q$(tail), $p + q = 1$. The information provided by the result 'head' is small: $I_h = -\ln p$ ($p \simeq 1$). On the other hand, the information provided by the infrequent result 'tail' is much larger: $I_t = -\ln q$ ($q \simeq 0$). The average information provided by both results is $\langle I \rangle = -p \ln p - q \ln q$ and $I_t > \langle I \rangle > I_h$. In the crystallographic case, we have I(negative triplet) $> \langle I \rangle > I$(positive triplet).

For single triplets, the information (3.4*b*) reduces to

$$I = \frac{1}{2(1-U_{\mathbf{h}-\mathbf{k}}^2)}\{E_{\mathbf{h}}^2 + E_{\mathbf{k}}^2 - 2|E_{\mathbf{h}}U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}|s(\mathbf{hk})\}, \tag{7.1}$$

which takes the second-order approximation form (6.2) if we neglect $U_{\mathbf{h}-\mathbf{k}}^2$ in comparison with 1 in (7.1).

Next we shall consider the expected information $\langle I\rangle$ in the following four cases:

(i) We lack any information about the structure factors $E_{\mathbf{h}}$ and $E_{\mathbf{k}}$ but we know their correlation $\langle E_{\mathbf{h}}E_{\mathbf{k}}\rangle = U_{\mathbf{h}-\mathbf{k}}$. Then, the expression analogous to (5.1) for $P\bar{1}$ is in this case

$$\langle I\rangle = \tfrac{1}{2}m = 1. \tag{7.2}$$

(ii) We have experimental information about the moduli $|E_{\mathbf{h}}|$ and $|E_{\mathbf{k}}|$. Then the average of (7.1), assuming random atom distribution, gives

$$\langle I\rangle = \frac{1}{2(1-U_{\mathbf{h}-\mathbf{k}}^2)}\{E_{\mathbf{h}}^2 + E_{\mathbf{h}}^2 - 2|E_{\mathbf{h}}U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \times \tanh|E_{\mathbf{h}}U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}|/(1-U_{\mathbf{h}-\mathbf{k}}^2)\}, \tag{7.3}$$

where the relation

$$\langle s(\mathbf{hk})\rangle = \tanh\frac{|E_{\mathbf{h}}U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}|}{1-U_{\mathbf{h}-\mathbf{k}}^2}, \tag{7.4}$$

has been used (Tsoucaris, 1970).

(iii) We know $|E_{\mathbf{h}}|$ and $|E_{\mathbf{k}}|$ from diffraction data and we also have structural information about the rough orientation of the molecule in the cell (obtained, for this crystal, from the inspection of the Patterson map). The averaging of (7.1) constrained by this additional condition gives

$$\langle I\rangle = \frac{1}{2(1-U_{\mathbf{h}-\mathbf{k}}^2)}\{E_{\mathbf{h}}^2 + E_{\mathbf{k}} - 2|E_{\mathbf{h}}U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}| \times \tanh[\tfrac{4}{3}N^2|E_{\mathbf{h}}U_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}|(\xi_{\mathbf{hk}} + 1/N^2)]\}, \tag{7.5}$$

where $\xi_{\mathbf{hk}}$ is defined by Kroon & Krabbendam (1970) and depends only on the interatomic vectors for the known molecular orientation.

(iv) The same as (iii) but now the refined molecular orientation is known.

The results of the calculations for the 22 large triplet products of low order reported by Kroon & Krabbendam (1970) are in Table 1, where the triplets have been numbered according to the order in which they appear in this reference. For the purpose of comparison, the table includes the real value [(7.1)] and the minimum value [(7.1) with $s(\mathbf{hk}) = 1$] for the information content of the single triplet products.

From an inspection of Table 1, we can conclude the following.

(i) The real value for the information carried by a few triplet products is smaller than the expected value in the case of complete ignorance about the structure factors $E_{\mathbf{h}}$ and $E_{\mathbf{k}}$. This fact correlates with triplets $E_{-\mathbf{h}}E_{\mathbf{h}-\mathbf{k}}E_{\mathbf{k}}$ of small modulus and emphasizes the advantage of working with large triplet products of higher information content.

(ii) The inequalities $\langle I\rangle_{r(c).o.} > \langle I\rangle_{r.a.}$ for negative triplets and $\langle I\rangle_{r(c).o.} < \langle I\rangle_{r.a.}$ for positive triplet products reflect the fact that the introduction of *a priori* structural knowledge improves sign prediction (Kroon & Krabbendam, 1970). When the molecular orientation is taken into account, the probability of a minus sign for the triplet products in the first group of Table 1 increases (together with the associated expected information) with respect to the values that are obtained by assuming a random atom distribution. The reverse effect occurs with the triplet products in the second group of Table 1.

(iii) When we increase our structural knowledge starting with the random atom distribution hypothesis (from left to right along the last three columne of Table 1), the expected information, in general, gets closer to the real value.

Table 1. *Information content for large triplet products of low order in* 1,4-*cyclohexanedione*

Triplet	Real I	Minimum I	$\langle I\rangle_{r.a.}$	$\langle I\rangle_{r.o.}$	$\langle I\rangle_{c.o.}$
			Negative triplets		
3	2·64	1·78	2·04	2·48	2·39
4	0·97	0·71	0·82	0·81	0·86
9	0·65	0·40	0·50	0·50	0·55
11	1·10	0·72	0·87	0·95	0·95
17	1·45	1·02	1·19	1·22	1·26
18	1·81	1·38	1·55	1·53	1·56
20	1·04	0·66	0·81	0·95	0·91
21	2·46	1·28	1·56	2·43	2·45
22	1·50	0·70	0·94	1·22	1·23
Average	1·51	0·96	1·14	1·34	1·35
			Positive triplets		
1	1·96	1·96	2·14	1·99	1·96
2	1·84	1·84	2·03	2·05	2·04
5	1·23	1·23	1·47	1·23	1·25
6	0·71	0·71	0·82	0·83	0·84
7	0·74	0·74	0·90	0·90	0·88
10	0·68	0·68	0·90	1·02	0·78
12	0·99	0·99	1·16	1·20	1·19
13	1·93	1·93	2·19	1·93	1·94
14	1·05	1·05	1·19	1·19	1·17
15	3·99	3·99	4·05	3·99	3·99
16	1·72	1·72	2·02	1·72	1·75
19	0·79	0·79	0·94	0·95	0·95
Average	1·47	1·47	1·65	1·58	1·56

$\langle I\rangle_{r.a.}$: expected information obtained from equation (7.3);
$\langle I\rangle_{r.o.}$: expected information obtained from equation (7.5) with rough orientation;
$\langle I\rangle_{c.o.}$: expected information obtained from equation (7.5) with correct orientation.
The triplet 8 has been omitted because it was clearly in error in the reference cited in the text.

8. Concluding remarks

The potentiality of information theory with regard to the derivation of the distribution function $P(\mathbf{E})$ of several structure factors compatible with constraints that embody *a priori* structural knowledge is shown in § 2. Once the distribution function $P(\mathbf{E})$ is found and the moduli of the components of $\mathbf{E}$ are experimentally determined, subsequent phase prediction based on the minimum added information $\mathbf{E}^+ \mathbf{U}^{-1} \mathbf{E}$ [or, to some degree of approximation, of its expansion (3.4)] can equally well be obtained with the equivalent condition of maximum $P(\mathbf{E})$. However, the information-theory approach offers some potentially useful insights for crystal structure analysis, derived from appropriate quantitation of the information provided by diffraction data and available stereochemical knowledge:

(i) A crystal structure may be regarded as a particular atomic arrangement pertaining to a set of many possible configurations. Therefore, a conformational entropy might be defined for such a crystal (Hosoya & Tokonami, 1976; Gassmann, 1977). In order to solve the structure, this uncertainty must be removed by diffraction data and stereochemical information. This information, quantified by expressions similar to (4.3), should match the conformation entropy. Also, the improvement derived from additional knowledge may be estimated comparing the magnitude of combined diffraction data and stereochemical information [*cf.* (5.8)], as opposed to that provided by diffraction data alone under the assumption of random atom distribution [*cf.* (5.3)]. In this connection, the corresponding gain in information, expressed to the second order of approximation by (5.9), should be non-negative when computed in a real case, provided that : (*a*) a statistically significant number of contributing triplet products are included; (*b*) the correct stereochemical data are employed; (*c*) the effect of terms higher than second order in (4.3) may be neglected.

(ii) The information included in a set of structure factors phased by multisolution direct methods can be quantitated by expressions similar to (4.3) to any degree of approximation. Work is in progress to assess the possibility of using such a measure as a useful figure or merit to select the correct phase set.

The expansion (4.3) in conjunction with the rule of minimum added information provide a basis for a systematic and consistent method of dealing with phase relationships simultaneously involving triplets, quartets, *etc.* Examples of this procedure are given in § 4.

After this paper was submitted, it came to our attention that others (Britten & Collins, 1982; Narayan & Nityananda, 1982) have also recognized the point discussed in §§ 2 and 3 of the present work, namely the relation between maximum entropy and maximum of Karle–Hauptman determinants.

I thank Dr A. D. Podjarny for helpful discussions and suggestions. I am pleased to acknowledge the interest and continuous encouragement of Dr P. B. Sigler to whom I am grateful. I am indebted to Dr E. E. Castellano for his interest and constructive criticism. Thanks are extended to Dr P. J. Aymonino for kindly correcting the manuscript. I am grateful to Dr J. Kroon who stressed the value of equation (4.3) in predicting the phases of large quartet products as described in the first paragraph of § 4(iii). This work was partially supported by grants from the NSF (Int 78-21875) and the NIH (GM 22324).

APPENDIX

The relation (2.9) between the β matrix and the Karle–Hauptman U matrix

The derivation of (2.9) is similar to that found in statistical mechanics when calculating fluctuations of thermodynamic quantities (Landau & Lifshitz, 1969). For convenience, we will deal here with the $P1$ case.

To condense the notation, we shall set:

$$E_l = E_{\mathbf{h}_l} \quad \text{and} \quad U_{jl} = U_{\mathbf{h}_j - \mathbf{h}_l}$$
$$(i, j = 1, 2, \ldots, m).$$

We shall assume that the $\mathbf{U}$ matrix is non-singular.

Let us define the random variables X_k by the relations

$$X_k = \sum_{p=1}^{m} \beta_{pk} E_p^* \quad (k = 1, 2, \ldots, m), \tag{A1}$$

and calculate the averages $\langle E_r X_k \rangle$ using the distribution function $P(\mathbf{E})$ given by (2.8):

$$\langle E_r X_k \rangle = C \int E_r \left(\sum_{p=1}^{m} \beta_{pk} E_p^* \right) \times \exp\left(- \sum_{i,j=1}^{m} \beta_{ij} E_i^* E_j \right) \mathrm{d}^m \mathbf{E}. \tag{A2}$$

To this purpose we shall assume for the moment that the averages $\langle E_r \rangle$ are not equal to zero, but equal to certain non-null values E_{ro}. Then, by definition,

$$\langle E_r \rangle = C \int E_r \exp \left\{ - \sum_{i,j=1}^{m} \beta_{ij} (E_i^* - E_{io}^*) \times (E_j - E_{jo}) \right\} \mathrm{d}^m \mathbf{E} = E_{ro}. \tag{A3}$$

Differentiating (*A*3) with respect to E_{ko}† and then setting again to zero all averages $E_{1o}, E_{2o}, \ldots, E_{mo}$, we obtain δ_{rk} for the third member of (*A*3), while the integral becomes equal to the average (*A*2):

† In $P1$, E_{ko} and E_{ko}^* must be considered as independent variables.

$$\sum_{p=1}^{m} \beta_{pk}\langle E_r E_p^* \rangle = \delta_{rk}. \qquad (A4)$$

Multiplying both members of (A4) by β_{ks}^{-1} and summing over the index k, we obtain

$$\sum_{p=1}^{m}\left(\sum_{k=1}^{m} \beta_{pk}\beta_{ks}^{-1}\right)\langle E_r E_p^* \rangle = \langle E_r E_s^* \rangle = U_{rs} = \beta_{rs}^{-1}, \qquad (A5)$$

where the Sayre–Hughes equation (2.5) has been used. Therefore

$$\boldsymbol{\beta} = \mathbf{U}^{-1}, \qquad (A6)$$

which is (2.9) for the $P1$ case. The proof of the relation $\boldsymbol{\beta} = \frac{1}{2}\mathbf{U}^{-1}$ that holds in the $P\bar{1}$ case closely follows the same steps as in the $P1$ case.

References

BRILLOUIN, L. (1962). *Science and Information Theory*. New York: Academic Press.

BRITTEN, P. L. & COLLINS, D. M. (1982). *Acta Cryst.* A**38**, 129–132.

COCHRAN, W. & WOOLFSON, M. M. (1955). *Acta Cryst.* **8**, 1–12.

DIAMOND, R. (1963). *Acta Cryst.* **16**, 627–639.

GASSMANN, J. (1977). *Acta Cryst.* A**33**, 474–479.

GERMAIN, G., MAIN, P. & WOOLFSON, M. M. (1970). *Acta Cryst.* B**26**, 274–285.

HAUPTMAN, H. (1974). *Acta Cryst.* A**30**, 472–476.

HEINERMAN, J. J. L. (1977). *Acta Cryst.* A**33**, 100–106.

HEINERMAN, J. J. L., KRABBENDAM, H. & KROON, J. (1977). *Acta Cryst.* A**33**, 873–878.

HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136–143.

HOSOYA, S. & TOKONAMI, M. (1976). *Acta Cryst.* **23**, 18–25.

HUGHES, E. W. (1953). *Acta Cryst.* **6**, 871.

KARLE, J. & KARLE, I. L. (1966). *Acta Cryst.* **21**, 849–859.

KOENIG, D. F. (1972). *Acta Cryst.* A**28**, 55.

KOENIG, D. F. (1976). *Chem. Scr.* **9**, 14–17.

KROON, J. & KRABBENDAM, H. (1970). *Acta Cryst.* B**26**, 312–314.

LANDAU, L. D. & LIFSHITZ, E. M. (1969). *Statistical Physics*, Vol. 5, pp. 345–348. Oxford: Pergamon Press.

MAIN, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 97–105. Copenhagen: Munksgaard.

MISES, R. VON (1918). *Phys. Z.* **19**, 490–500.

MOSSEL, A. & ROMERS, C. (1964). *Acta Cryst.* **17**, 1217–1223.

NARAYAN, R. & NITYANANDA, R. (1982). *Acta Cryst.* A**38**, 122–128.

PIRO, O. E. (1977). Thesis, La Plata, Argentina.

PIRO, O. E. & PODJARNY, A. D. (1978). 36th Annual Pittsburgh Diffraction Conference, Abstract P-12, p. 35.

PIRO, O. E. & PODJARNY, A. D. (1979). Am. Crystallogr. Assoc. Winter Meet., Abstract J6, p. 76.

POST, B. (1979). *Acta Cryst.* A**35**, 17–21.

RANGO, C. DE, TSOUCARIS, G. & ZELWER, C. (1974). *Acta Cryst.* A**30**, 342–353.

SAYRE, D. (1952). *Acta Cryst.* **5**, 60–65.

SCHENK, H. (1973). *Acta Cryst.* A**29**, 77–82.

SCHENK, H. (1974). *Acta Cryst.* A**30**, 477–481.

SHANNON, C. E. & WEAVER, W. (1949). *The Mathematical Theory of Communication*. Urbana: Univ. of Illinois Press.

TSOUCARIS, G. (1970). *Acta Cryst.* A**26**, 492–499.

Acta Cryst. (1983). A**39**, 68–74

Phase Extension and Refinement by Density Modification in Protein Crystallography

BY E. CANNILLO, R. OBERTI AND L. UNGARETTI

CNR Centro di Studio per la Cristallografia Strutturale, c/o Istituto di Mineralogia, via Bassi 4, 27100 *Pavia, Italy*

(*Received* 22 *February* 1982; *accepted* 6 *July* 1982)

Abstract

A new procedure of phase extension and refinement *via* electron density modification applicable to low-resolution protein crystal structures is described. The *Sperm Whale* myoglobin structure has been used as a working molecule. The procedure of phase extension has firstly been tested starting from a set of calculated phases at 4 Å resolution; the mean phase error obtained for the 9000 strongest reflections from 4 to 1·8 Å was 39°; subsequently a mean phase error of 30° was spread into the low-resolution set and a phase refinement and extension procedure was carried out to 1·8 Å resolution. The final mean phase errors of the 1184 low-resolution model and of the 4816 strongest reflections within 1·8 Å were 22 and 50° respectively. The map calculated with this final set of reflections approaches in quality and details the map calculated with the 12 658 phases from the refined coordinates.

Introduction

A crucial step in modern protein crystallography is the calculation of a good quality electron density map of medium-to-high resolution, suitable for model building and/or least-squares refinement.

Multiple isomorphous replacement methods (MIR) very often do not achieve this goal: the crystal-

0567-7394/83/010068-07$01.50