Acta Cryst. (1982). A**38**, 401–407

A New Determinantal Theory for Solving the Phase Problem Using Stereochemical Information

BY A. D. PODJARNY* AND C. FAERMAN†

Laboratorio de Rayos X, UNLP, CC 67, 1900 *La Plata, Argentina*

(*Received* 7 *May* 1981; *accepted* 29 *September* 1981)

Abstract

A priori structural information is incorporated into determinantal phasing techniques to improve phase-prediction accuracy in resolution ranges where atomic or isotropic group-scatterer assumptions are not valid. For this purpose a conditional joint probability distribution to triplet order for any set of normalized structure factors of space groups $P1$ and $P\bar{1}$ is derived. The covariance of two normalized structure factors from the original set is calculated. A more general conditional joint probability distribution, involving covariance matrices of any order, is further derived. Numerical tests are performed employing ideal models consisting of several atomic groups of known stereochemistry but with random positions and orientations. The results indicate that the inclusion of stereochemical information improves the accuracy of phase prediction. The relative merit of this strategy in either one of or both normalization and covariance calculations for different resolutions is discussed.

Introduction

Determinantal direct methods have previously been used to improve high-resolution phases for macromolecules (de Rango, Mauguen & Tsoucaris, 1975; Podjarny, Yonath & Traub, 1976; Podjarny & Yonath, 1977). These techniques are useful when the available MIR data approach atomic resolution. In this instance several other approaches are also available, such as Sayre's formula and constrained–restrained refinement (Sayre, 1974; Sussman, Holbrook, Church & Kim, 1977). These methods not only are effective but also, in the case of the constrained–restrained refinement, make optimum use of available stereochemical information, provided that a rough model is available.

However, in a large number of cases, intensities and/or MIR phases do not reach atomic resolution. Therefore, it is convenient to introduce stereochemical information at an early stage of the phasing procedure, before an approximate model is obtained. Such an approach has been introduced in the phasing of small-molecule structure factors by Main (1976), and expressed in stricter theoretical terms by Heinerman (1977). This theory introduces stereochemical information in the prediction of the phase of a triple product of normalized structure factors. The information is supplied in the form of the stereochemistry of rigid partial structures, which can have (1) fixed orientation and position, (2) fixed orientation and random position or (3) random orientation and position. For the case of fixed orientation and random position, tests performed by Main (1976) showed significant improvements in the accuracy of phase prediction.

In the case of macromolecules, the tangent formula (Karle & Hauptman, 1956) does not have, in general, enough predictive power [for a discussion of this point see Mauguen (1979) and references therein]. It is therefore of interest to discover whether determinants, which involve higher-order n-tuplets, could be used, in conjunction with stereochemical information, at low and medium resolutions. In addition, one has commonly to deal here with partial structures of random position and orientation but known stereochemistry. The partial structures or atomic groups are, for example, the monomers of the biopolymer (peptides in the case of proteins, nucleotides in the case of nucleic acids).

For very-low-resolution data, these atomic groups can be considered as isotropic 'atoms' and determinants can be used with slight modifications. Such an approach was employed by Podjarny, Schevitz & Sigler (1981) for the case of tRNA. However, they have observed that the assumption of isotropic 'atoms' breaks down rapidly with increasing resolution.

Therefore, we have extended the theory for triplet-phase prediction developed by Heinerman (1977) to covariance-matrix methods of phasing. Our purpose in this work is to add, to the experimental information contained in medium-resolution macromolecular data, stereochemical information generally known *a priori*. Our experience shows that, in resolution ranges where the assumption of 'isotropic group scatterers' (Podjarny, Schevitz & Sigler, 1981; Podjarny & Yonath, 1977) is valid, it is possible to extend phases with only a

* Member of CONICET.

† Supported by a fellowship from CICPBA.

0567-7394/82/040401-07$01.00

change in data normalization. However, work now in progress shows that in the resolution range where the 'isotropic group scatterer' assumption is not valid, phase-prediction accuracy decreases even with properly normalized data. Therefore, we tried to include the internal group stereochemistry in the whole theory. For this purpose, we choose an arbitrary set of m reciprocal vectors (the generating ones) and take the corresponding E's as the components of a random vector of dimension m (the vector of generating E's). We then calculate to the triplet order the conditional probability distribution of this generating set of structure factors when all the moduli and phases of the possible triplets involving two generating reflections are known. This distribution is used to calculate the covariances, also to triplet order, which are arranged in a covariance matrix. Using the central-limit theorem, a more general probability distribution is obtained, which includes all n-tuplets up to order $m + 1$, functionally related to the triplets in a way similar to that in which the moments of a unidimensional Gaussian distribution are related to the variance.

This approach is central to the development of probability distributions using covariance matrices (Tsoucaris, 1970; Castellano, Podjarny & Navaza, 1973). The central-limit theorem is particularly suitable in the case of macromolecules since the number of original random variables, *i.e.* the atomic positions, is quite large.

The effect of the present theory is to modify the normalization and the covariances according to the stereochemistry. Results are described below for several test cases.

Notation

We follow broadly the notation of Heinerman (1977) and Heinerman, Krabbendam & Kroon (1979). $g_k(\mathbf{h}) = \sum_{L=1}^{n_k} f_K^L(\mathbf{h}) \exp[2\pi i \mathbf{h}.(\mathbf{r}_k^L - \mathbf{r}_k)]$ is the group scattering factor, $f_k^L(\mathbf{h})$ is the atomic scattering factor, $\mathbf{r}_k$ is the center of mass for the kth group, $\mathbf{r}_k^L$ is an atomic coordinate with respect to this center, and n_k is the number of atoms in this group. $\mathbf{h}_{oi_1}$ is a 'generating' reciprocal vector; ($\mathbf{h}_{00} = \mathbf{0}$).

$$\prod' = \prod_{\substack{i_1,i_2=0\\ i_1<i_2}}^{m}; \quad \sum' = \sum_{\substack{i_1,i_2=0\\ i_1<i_2}}^{m}; \quad \prod'' = \prod_{\substack{i_4,i_5=0\\ i_4<i_5\\ i_4i_5\neq i_1i_2\\ \neq i_2i_3\\ \neq i_1i_3}}^{m}; \quad \sum'' = \sum_{\substack{i_1,i_2,i_3=0\\ i_1<i_2<i_3}}^{m}.$$

FM(φ^p) is the figure of merit for the predicted value of phase φ.

$$\prod''' = \prod_{\substack{i_1=0\\ i_2=1\\ i_1\neq i_2}}^{m}; \quad \sum''' = \sum_{\substack{i_1=0\\ i_2=1\\ i_1\neq i_2}}^{m} \quad \sum^{\mathrm{iv}} = \sum_{\substack{i_2=1\\ i_3=1\\ i_2<i_3}}^{m}.$$

Derivation of the conditional probability for m normalized structure factors using stereochemical information for space groups $P1$ and $P\bar{1}$

(1) *Space group P*1

The normalized structure factors are

$$E_{\mathbf{h}_{i_1i_2}} = \sum_{j=1}^{p} g_j(\mathbf{h}_{i_1i_2}) \exp(2\pi i \mathbf{h}_{i_1i_2}.\mathbf{r}_j)/\langle |F_{\mathbf{h}_{i_1i_2}}|^2\rangle_{\text{p.r.v.}}^{1/2}$$

where $g_j(\mathbf{h}_{i_1i_2})$ is the atomic scattering factor for $1 \le j \le p_1$ and the group scattering factor for $p_1 + 1 \le j \le p$ and where $\langle |F_{\mathbf{h}_{i_1i_2}}|^2\rangle$ is the average of $|F_{\mathbf{h}_{i_1i_2}}|^2$ over the following primitive random variables: (*a*) $1 \le j \le p_1$, the atomic position vectors; (*b*) $p_1 + 1 \le j \le p_2$, the position vectors and orientational parameters of the randomly positioned and orientated groups, (*c*) $p_2 + 1 \le j \le p$, the position vectors of the groups with known orientations. (Note that p is the total number of groups, considering a single independent atom as a group.)

Alternatively, we can express $E_{\mathbf{h}_{i_1i_2}}$ as follows:

$$E_{\mathbf{h}_{i_1i_2}} = \sum_{j=1}^{p} u_j(\mathbf{h}_{i_1i_2}) \exp[2\pi i \mathbf{h}_{i_1i_2}.\mathbf{r}_j + i\beta_j(\mathbf{h}_{i_1i_2})]$$

where we have used the following definition:

$$u_j(\mathbf{h}_{i_1i_2}) \exp[i\beta_j(\mathbf{h}_{i_1i_2})] = g_j(\mathbf{h}_{i_1i_2})/\langle |F_{\mathbf{h}_{i_1i_2}}|^2\rangle_{\text{p.r.v.}}^{1/2}$$

and $u_j(\mathbf{h}_{i_1i_2})$ is a real positive number.

For a detailed calculation of $\langle |F_{\mathbf{h}_{i_1i_2}}|^2\rangle_{\text{p.r.v.}}$ see Main (1976) and Heinerman (1977). These results were used in our theoretical calculation of intensity curves.

Operating in polar coordinates, we define:

$$E_{\mathbf{h}_{i_1i_2}} = R_{i_1i_2} \exp(i\varphi_{i_1i_2}) \quad (i_1, i_2 = 0, \ldots, m)$$

and $\mathbf{h}_{i_1i_2} = \mathbf{h}_{0i_2} - \mathbf{h}_{0i_1}$ with $i_l < i_2$. We used the method of the characteristic function to obtain the joint probability distribution as a function of the variables $R_{i_1i_2}$, $\varphi_{i_1i_2}$ $(i_1, i_2 = 0, \ldots, m)$ which is given by:

$$P(R_{01}, \ldots, R_{i_1i_2}, \ldots; \varphi_{01}, \ldots, \varphi_{i_1i_2}, \ldots)$$

$$= \frac{1}{(2\pi)^{m(m+1)}} \prod^{m\prime} R_{i_1i_2}$$

$$\times \underbrace{\int_0^{\infty}\int_0^{2\pi} \cdots \int_0^{\infty}\int_0^{2\pi}}_{\substack{m(m+1)/2\\ \text{double integrals}}} \exp\left[-i \sum^{m\prime} R_{i_1i_2}\rho_{i_1i_2}\cos(\varphi_{i_1i_2} - \theta_{i_1i_2})\right]$$

$$\times \prod^{m\prime} Q(\rho_{01}, \ldots, \rho_{i_1i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1i_2}, \ldots)$$

$$\times \rho_{i_1i_2}\, \mathrm{d}\rho_{i_1i_2}\, \mathrm{d}\theta_{i_1i_2}$$

where

$$Q(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots)$$
$$= \left\langle \prod_{j=1}^{p} \exp\left\{ i \sum_{}^{m\prime} u_j(\mathbf{h}_{i_1 i_2}) \rho_{i_1 i_2} \cos[2\pi \mathbf{h}_{i_1 i_2} . \mathbf{r}_j + \beta_j(\mathbf{h}_{i_1 i_2}) - \theta_{i_1 i_2}] \right\} \right\rangle_{\text{p.r.v.}}.$$

If we assume that the p different groups are independent, we can then express the characteristic function as a product of the characteristic functions of the different groups:

$$Q(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots)$$
$$= \prod_{j=1}^{p} q_j(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots)$$

where

$$q_j(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots)$$
$$= \left\langle \exp\left\{ i \sum_{}^{m\prime} u_j(\mathbf{h}_{i_1 i_2}) \rho_{i_1 i_2} \cos[2\pi \mathbf{h}_{i_1 i_2} . \mathbf{r}_j + \beta_j(\mathbf{h}_{i_1 i_2}) - \theta_{i_1 i_2}] \right\} \right\rangle_{\text{p.r.v.}j}.$$

Developing the exponential in terms of Bessel functions, up to triplet order only, we obtain (see Appendix I*a*):

$$Q(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots)$$
$$\simeq \exp\left[-\tfrac{1}{4} \sum_{}^{m\prime} \rho_{i_1 i_2}^2 - \tfrac{1}{4} i \sum_{}^{m\prime\prime} \times Q_{i_1 i_2, i_2 i_3, i_1 i_3} \rho_{i_1 i_2} \rho_{i_2 i_3} \rho_{i_1 i_3} \times \cos(\theta_{i_1 i_2} + \theta_{i_2 i_3} - \theta_{i_1 i_3} - q_{i_1 i_2, i_2 i_3, i_1 i_3}) \right]$$

where $Q_{i_1 i_2, i_2 i_3, i_1 i_3} \exp(i q_{i_1 i_2, i_2 i_3, i_1 i_3})$ is given in Appendix I(*a*).

To integrate we used the procedure suggested by Heinerman, Krabbendam & Kroon (1979) and finally obtained:

$$P(R_{01}, \ldots, R_{i_1 i_2}, \ldots; \varphi_{01}, \ldots, \varphi_{i_1 i_2}, \ldots)$$
$$= \frac{1}{\pi^{m(m+1)/2}} \prod_{}^{m\prime} R_{i_1 i_2} \exp\left[- \sum_{}^{m\prime} R_{i_1 i_2}^2 + 2 \sum_{}^{m\prime\prime} Q_{i_1 i_2, i_2 i_3, i_1 i_3} R_{i_1 i_2} R_{i_2 i_3} R_{i_1 i_3} \times \cos(\varphi_{i_1 i_2} + \varphi_{i_2 i_3} - \varphi_{i_1 i_3} - q_{i_1 i_2, i_2 i_3, i_1 i_3}) \right].$$

From this general distribution we obtained the conditional probability distribution up to triplet order:

$$P_{\text{cond.}}[R_{01}, \ldots, R_{0m}; \varphi_{01}, \ldots, \varphi_{0m} / \ldots R_{i_1 i_2} \ldots; \ldots \varphi_{i_1 i_2} \ldots (i_1 \neq 0, i_2 \neq 0; i_1 < i_2)]$$
$$= \frac{P(R_{01}, \ldots, R_{i_1 i_2}, \ldots; \varphi_{01}, \ldots, \varphi_{i_1 i_2}, \ldots)}{P_{\text{marg}}[\ldots R_{i_1 i_2} \ldots; \ldots \varphi_{i_1 i_2} \ldots (i_1 \neq 0, i_2 \neq 0; i_1 < i_2)]}$$
$$= \frac{1}{\pi^m} \prod_{i_1=1}^{m} R_{0 i_1} \exp\left[- \sum_{i_1=1}^{m} R_{0 i_1}^2 + \sum_{\substack{i_1, i_2 = 1 \\ i_1 < i_2}}^{m} 2 Q_{0 i_1, i_1 i_2, 0 i_2} R_{0 i_1} R_{i_1 i_2} R_{0 i_2} \times \cos(\varphi_{0 i_1} + \varphi_{i_1 i_2} - \varphi_{0 i_2} - q_{0 i_1, i_1 i_2, 0 i_2}) \right].$$

This allowed us to calculate conditional covariances, also up to triplet order, the expression of which is as follows:

$$\langle E_{0 i_1} E_{0 i_2}^* \rangle \ldots E_{i_1 i_2} \ldots (i_1 \neq 0, i_1 < i_2)$$
$$= Q_{0 i_1, i_1 i_2, 0 i_2} \exp(+ i q_{0 i_1, i_1 i_2, 0 i_2}) E_{i_1 i_2}^*. \qquad (1)$$

Note that $\langle E_{0 i_2} E_{0 i_1}^* \rangle = \langle E_{0 i_1} E_{0 i_2}^* \rangle^*$; $\langle E_{0 i_1} E_{0 i_1}^* \rangle = 1$. To obtain higher-order relationships dependent only upon the covariances we can follow the usual procedure of invoking the central-limit theorem, which leads us to the following conditional probability distribution:

$$P(E_{01}, \ldots, E_{0m} / P_{11}, \ldots) = \frac{1}{\pi^m} \frac{1}{D_m} \exp(-Q_m)$$

(Tsoucaris, 1970)

where

$$Q_m = \sum_{i_1, i_2 = 1}^{m} E_{0 i_1}^* D_{i_1 i_2} E_{0 i_2},$$

D_m is the determinant of the correlation matrix and $D_{i_1 i_2}$ is an element of the inverse correlation matrix. From this formula we can calculate the statistical regression of any generating reflection, say E_{0m}, upon all the others. We obtain for the predicted value of E_{0m}:

$$E_{0m}^p = \sum_{i_1, i_2 = 1}^{m-1} D_{i_1 i_2} E_{0 i_2} = |E_{0m}^p| \exp(i \varphi_{0m}^p)$$

with the corresponding figure of merit for the predicted value of the phase φ_{0m}:

$$\text{FM}(\varphi_{0m}^p) = I_1(B)/I_0(B)$$

where $B = 2|E_{0m}||E_{0m}^p|$; I_1 and I_0 are modified Bessel functions.

(2) *Space group* $P\bar{1}$

The structure factor expression is:

$$E_{\mathbf{h}_{i_1 i_2}} = \sum_{j=1}^{p/2} 2u_j(\mathbf{h}_{i_1 i_2}) \cos[2\pi \mathbf{h}_{i_1 i_2}.\mathbf{r}_j + \beta_j(\mathbf{h}_{i_1 i_2})]$$

where

$$u_j(\mathbf{h}_{i_1 i_2}) = \frac{|g_j(\mathbf{h}_{i_1 i_2})|}{\langle |F_{\mathbf{h}_{i_1 i_2}}|^2 \rangle^{1/2}_{\text{p.r.v.}}}.$$

Note that, if $F(2\mathbf{h}_{i_1 i_2})$ is not known, $\langle |F_{\mathbf{h}_{i_1 i_2}}|^2 \rangle^{1/2}_{\text{p.r.v.}} = \sum_{j=1}^{p/2} |g_j(\mathbf{h}_{i_1 i_2})|^2$ (Main, 1976). Now the definition of $\mathbf{h}_{i_1 i_2}$ for $P\bar{1}$ is:

$$\mathbf{h}_{i_1 i_2} = \mathbf{h}_{0 i_2} - \mathbf{h}_{0 i_1} \quad \text{if } i_1 < i_2$$
$$\mathbf{h}_{i_1 i_2} = \mathbf{h}_{0 i_1} + h_{0 i_2} \quad \text{if } i_1 > i_2.$$

The primitive random variables are : (*a*) $1 \le j \le p_1/2$, the atomic position vectors; (*b*) $p_1/2 + 1 \le j \le p_2/2$, the position vectors and orientational parameters of the randomly positioned and orientated groups; (*c*) $p_2/2 + 1 \le j \le p/2$, the position vectors of the groups with known orientation.

Using again the method of the characteristic function, we obtained the joint probability distribution, which expression is as follows:

$$P(E_{01}, \ldots, E_{i_1 i_2}, \ldots) = \frac{1}{(2\pi)^{m^2}} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{m^2 \text{ integrals}} \exp\left(-i \sum^{m}{}''' E_{i_1 i_2} \rho_{i_1 i_2}\right) \times Q(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots) \prod^{m}{}''' \mathrm{d}\rho_{i_1 i_2}$$

where

$$Q(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots) = \left\langle \prod_{j=1}^{p/2} \exp\left(i \sum^{m}{}''' \{2u_j(\mathbf{h}_{i_1 i_2}) \rho_{i_1 i_2} \times \cos[2\pi \mathbf{h}_{i_1 i_2}.\mathbf{r}_j + \beta_j(\mathbf{h}_{i_1 i_2})]\}\right) \right\rangle_{\text{p.r.v.}}.$$

Again, if the groups are independent:

$$Q(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots) = \prod_{j=1}^{p/2} \left\langle \exp\left(i \sum^{m}{}''' \{2u_j(\mathbf{h}_{i_1 i_2}) \rho_{i_1 i_2} \times \cos[2\pi \mathbf{h}_{i_1 i_2}.\mathbf{r}_j + \beta_j(\mathbf{h}_{i_1 i_2})]\}\right) \right\rangle_{\text{p.r.v.}j} = \prod_{j=1}^{p/2} q_j(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots).$$

Developing the exponential as a power series of Bessel functions we obtained (see Appendix I*b*):

$$Q(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots) \simeq \exp\left\{ -\tfrac{1}{2} \sum^{m}{}''' \rho^2_{i_1 i_2} - 2i \sum^{m}{}^{\text{iv}} [(Q_{0i_2, i_2 i_3, 0i_3} \rho_{0i_2} \rho_{i_2 i_3} \rho_{0i_3} \cos q_{0i_2, i_2 i_3, 0i_3}) + (Q_{0i_2, i_3 i_2, 0i_3} \rho_{0i_2} \rho_{i_3 i_2} \rho_{0i_3} \cos q_{0i_2, i_3 i_2, 0i_3})]\right\}.$$

After integrating, we finally obtained:

$$P(E_{01}, \ldots, E_{i_1 i_2}, \ldots) = \frac{1}{(\sqrt{2\pi})^{m^2}} \exp\left[-\tfrac{1}{2} \sum^{m}{}''' E^2_{i_1 i_2} + 2 \sum^{m}{}^{\text{iv}} (Q_{0i_2, i_2 i_3, 0i_3} E_{0i_2} E_{i_2 i_3} E_{0i_3} \cos q_{0i_2, i_2 i_3, 0i_3} + Q_{0i_2, i_3 i_2, 0i_3} E_{0i_2} E_{i_3 i_2} E_{0i_3} \cos q_{0i_2, i_3 i_2, 0i_3})\right].$$

As for space group $P1$, we obtained from the joint probability distribution the conditional one up to triplet order, which is given by:

$$P_{\text{cond.}}[E_{01}, \ldots, E_{0m} / \ldots E_{i_1 i_2} \ldots (i_1 \neq 0, i_2 \neq 0; i_1 \neq i_2)] = \frac{1}{(\sqrt{2\pi})^m} \exp\left[-\tfrac{1}{2} \sum_{i_1=1}^{m} E^2_{0i_1} + 2 \sum^{m}{}^{\text{iv}} (Q_{0i_2, i_2 i_3, 0i_3} E_{0i_2} E_{i_2 i_3} E_{0i_3} \cos q_{0i_2, i_2 i_3, 0i_3} + Q_{0i_2, i_3 i_2, 0i_3} E_{0i_2} E_{i_3 i_2} E_{0i_3} \cos q_{0i_2, i_3 i_2, 0i_3})\right].$$

According to the same criteria used for $P1$, we calculated the conditional covariances and invoking the central-limit theorem obtained finally another conditional probability. They are respectively given by:

$$\langle E_{0i_1} E_{0i_2} \rangle \ldots E_{i_1 i_2} \ldots; \ldots E_{i_2 i_1} \ldots (i_1 \neq 0, i_1 < i_2) = 2(E_{i_1 i_2} Q_{0i_1, i_1 i_2, 0i_2} \cos q_{0i_1, i_1 i_2, 0i_2} + E_{i_2 i_1} Q_{0i_1, i_2 i_1, 0i_2} \cos q_{0i_1, i_2 i_1, 0i_2}). \quad (2)$$

$$P(E_{01}, \ldots, E_{0m} / P_{11}, \ldots) = \frac{1}{(\sqrt{2\pi})^m} \frac{1}{D_m^{1/2}} \exp(-\tfrac{1}{2} Q_m)$$

(Tsoucaris, 1970).

From this formula the predicted value of a single phase is calculated as in the case of space group $P1$.

Results on test models

To test the theory developed above, several structural models were simulated. As the main interest of this work is future applications to biopolymers, the models consisted of several repetitions of the same group. The

resolution used was such that atoms were not fully resolved. This is meant to simulate cases where electron density maps would not be completely interpretable by standard procedures and some improvement of the image is necessary.

Groups are widely spaced and randomly positioned. They do not represent a real case, since the data are ideal. Instead, they provide a numerical test of the theory and assessment of the influence of stereochemical information on phase prediction.

The symmetry used was $P1$ and $P\bar{1}$. The resolutions used were 3 and 1·9 Å. A typical interatomic distance was 1·5 Å. The cell dimensions were 15 × 15 × 15 Å.

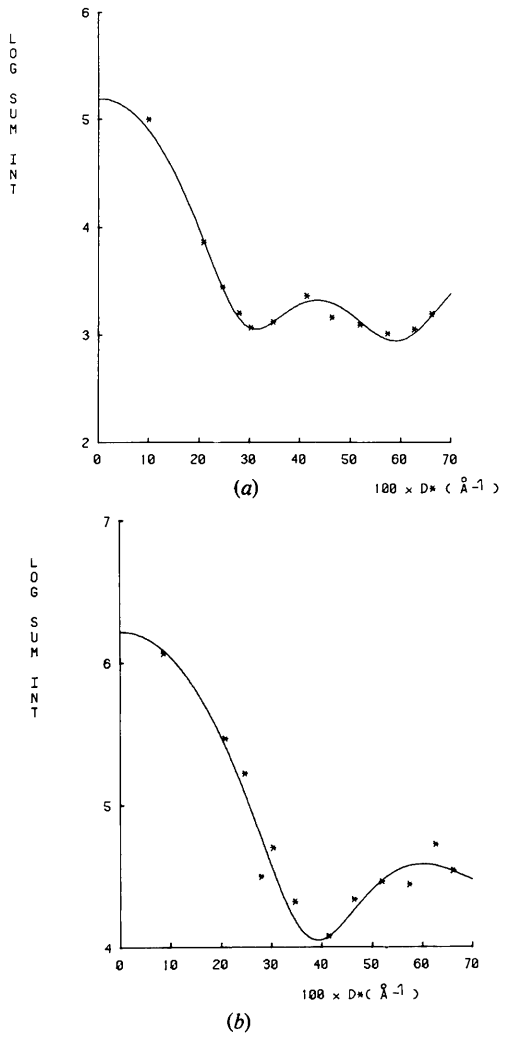

Fig. 1. (*a*) Intensity curve for hexagonal groups. The hexagon side length was 1·5 Å. The full curve was calculated from the Debye formula for rotationally and translationally averaged groups (see Main, 1976). The points were obtained from the calculated structure factors. (*b*) Same as (*a*), for tetrahedral groups with a center atom. The tetrahedron side length was 2·13 Å, and the closest interatomic distance was 1·3 Å.

The mean intergroup distance in the asymmetric unit was 8 Å. The group geometries were hexagonal and tetrahedral (with a center atom).

Ten groups were repeated per asymmetric unit by random translations and rotations. Structure factors were calculated and normalized, either with or without stereochemical information.

To normalize according to group geometry, the intensity curve was calculated from interatomic distances and compared with the observed values of the mean intensity. Such curves are shown in Fig. 1. Note that similar curves have been observed previously with experimental macromolecular data (French & Wilson, 1978; Podjarny *et al.*, 1981).

Karle–Hauptman and Goedkoop matrices were built using the largest structure factors as generating reflections. Out of several models tested, the case of a 5 × 5 matrix with a tetrahedral motif was selected, for both $P1$ and $P\bar{1}$. The resolutions used were 3 and 1·9 Å. For the latter resolution, the generating reflections were selected within the 3 Å sphere to simulate a dependence on low-resolution, high-figure-of-merit phases.

Single phases were predicted each time, following a previously reported technique (Podjarny *et al.*, 1976), in three different ways: (1) disregarding group geometry; (2) using group geometry for normalization only; (3) using group geometry for both normalization and covariance calculation purposes, according to formulae (1) and (2). Thus, it was possible to measure the relative influence of stereochemical information in normalization and covariance calculations. This is significant because the introduction of group geometries in the calculation of covariances for the case of

Table 1. *Root mean square error for phase extension* (°)

Case	S.G.	Res.	N. Gr.	RMS1	RMS2	RMS3	RMS4
1	$P1$	3	80	61	31 (129)	36	21
2	$P1$	3	20	59	27 (106)	24	18
3	$P1$	3	20	59	25 (97)	24	18
1	$P1$	1·9	80	65	32 (330)	36	25
2	$P1$	1·9	20	66	44 (505)	38	26
3	$P1$	1·9	20	65	28 (346)	36	25
1	$P\bar{1}$	3	80	72	26 (97)	23	13
2	$P\bar{1}$	3	20	68	0 (80)	23	4
3	$P\bar{1}$	3	20	69	0 (76)	23	5
1	$P\bar{1}$	1·9	80	86	24 (270)	40	15
2	$P\bar{1}$	1·9	20	90	40 (387)	46	17
3	$P\bar{1}$	1·9	20	87	19 (267)	36	13

'Case' is as described in the text, 'S.G.' is space group, 'Res.' is resolution in Å, 'N. Gr.' is the number of groups (atoms for case 1), RMS1 is the overall error (252 reflections for 3 Å; 1054 reflections for 1·9 Å), RMS2 is the r.m.s. error for $B > 1$ (number of reflections), RMS3 is the r.m.s. error for the largest F's (30 F's in $P1$, 3 Å; 60 F's in $P\bar{1}$, 3 Å; 200 F's in 1·9 Å, $P1$ and $P\bar{1}$), RMS4 is the same as RMS3, but weighted by B. Note that RMS4 is related to the correlation coefficient that links the original density with the predicted one (see Appendix II).

random translation and rotation requires lengthy computations. Table 1 lists the results of these predictions.

Conclusions

Although the results described above correspond to ideal cases, they illustrate how the introduction of stereochemical information decreases r.m.s. error for the same reflections, and that the number of groups needed is considerably less than the number of atoms. This had already been pointed out (Podjarny & Yonath, 1977) and in real cases permits the use of much smaller matrices.

The error for $B > 1$ (RMS2) is essentially a measure of the self consistency of the theory, whereas the overall error (RMS1) and the errors for the largest F's (RMS3 and RMS4) are a measure of its real usefulness. In general, it is observed that the introduction of stereochemical information is always useful if used for both normalization and covariance calculations.

Using RMS3 as a measure of phase-prediction accuracy, we can separate the results shown in Table 1 in two main groups:

(1) Improvement of phase-prediction accuracy, when normalization is performed according to group geometry, but without further improvement of it when stereochemistry is included also in covariance calculations. This is the case for 3 Å resolution, both in $P1$ and $P\bar{1}$. For space group $P1$, the effect is more evident.

(2) Improvement of phase-prediction accuracy when stereochemical information is used in both normalization and covariance calculations. However, a decrease of the accuracy is observed when group geometries are used in normalization only. This happens at 1·9 Å resolution, both in $P1$ and $P\bar{1}$.

It would seem that the effect of using group geometries only in normalization depends on the resolution of the data. It should be noted that other measures of error give much smaller differences than RMS3, but in the same sense. As the largest F's influence the map strongly, RMS3 seems to be a reasonable choice, but only practical applications can gauge the real usefulness of the method.

The improvements in phase extension that we obtained were smaller than those reported by Main (1976) using the tangent formula with groups of known orientation. This is probably due to the decrease in structural information by rotational averaging. However, in real macromolecular cases at medium resolution the orientation of known groups is not known, and techniques developed for atomic resolution do not work. Therefore, the theory hereby developed seems a viable alternative to improve phase-prediction accuracy. This might be of special significance in those medium-resolution cases where normalization only deteriorates phase prediction. Work is in progress with such cases, nucleic acids in particular, in order to assess the potential relevance of the phase improvements reported herein to the quality of the electron density image.

APPENDIX I

(*a*) *Space group P*1

If $p_1 + 1 \le j \le p_2$ then:

$$q_j(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots, \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots) = \left\langle \exp\left\{ i \sum_{}^{m\prime} u_j(\mathbf{h}_{i_1 i_2}) \rho_{i_1 i_2} \cos[2\pi \mathbf{h}_{i_1 i_2} \cdot \mathbf{r}_j + \beta_j(\mathbf{h}_{i_1 i_2}) - \theta_{i_1 i_2}] \right\} \right\rangle_{\mathbf{r}_j, \text{orient.}}.$$

Expanding the exponential in power series of Bessel functions according to formulae used by Heinerman (1977) and references therein, and considering terms up to triplet order, we finally obtained for $p_1 + 1 \le j \le p_2$

$$q_j(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots) \simeq 1 - \tfrac{1}{4} \sum_{}^{m\prime} \langle u_j^2(\mathbf{h}_{i_1 i_2}) \rangle_{\text{orient.}} \rho_{i_1 i_2}^2 - \tfrac{1}{4} i \sum_{}^{m\prime\prime} \langle u_j(\mathbf{h}_{i_1 i_2}) u_j(\mathbf{h}_{i_2 i_3}) u_j(\mathbf{h}_{i_1 i_3}) \rho_{i_1 i_2} \rho_{i_2 i_3} \rho_{i_1 i_3} \times \cos\{\beta_j(\mathbf{h}_{i_1 i_2}) + \beta_j(\mathbf{h}_{i_2 i_3}) - \beta_j(\mathbf{h}_{i_1 i_3}) - \theta_{i_1 i_2} - \theta_{i_2 i_3} + \theta_{i_1 i_3}\} \rangle_{\text{orient.}}.$$

From this expression the corresponding ones for $1 \le j \le p_1$ and $p_2 + 1 \le j \le p$ can be derived noting that for both $1 \le j \le p_1$ and $p_2 + 1 \le j \le p$ there is no orientational average and that for $1 \le j \le p_1$ $\beta_j(\mathbf{h}) = 0$. We finally obtained:

$$\prod_{j=1}^{p} q_j(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots; \theta_{01}, \ldots, \theta_{i_1 i_2}, \ldots) \simeq \exp\left[-\tfrac{1}{4} \sum_{}^{m\prime} \rho_{i_1 i_2}^2 - \tfrac{1}{4} i \sum_{}^{m\prime\prime} Q_{i_1 i_2, i_2 i_3, i_1 i_3} \rho_{i_1 i_2} \rho_{i_2 i_3} \rho_{i_1 i_3} \times \cos(\theta_{i_1 i_2} + \theta_{i_2 i_3} - \theta_{i_1 i_3} - q_{i_1 i_2, i_2 i_3, i_1 i_3}) \right],$$

where

$$Q_{i_1 i_2, i_2 i_3, i_1 i_3} \exp(i q_{i_1 i_2, i_2 i_3, i_1 i_3}) = \sum_{j=1}^{p_1} t_j(\mathbf{h}_{i_1 i_2}) t_j(\mathbf{h}_{i_2 i_3}) t_j^*(\mathbf{h}_{i_1 i_3}) + \sum_{j=p_1+1}^{p_2} \langle t_j(\mathbf{h}_{i_1 i_2}) t_j(\mathbf{h}_{i_2 i_3}) t_j^*(\mathbf{h}_{i_1 i_3}) \rangle_{\text{orient.}} + \sum_{j=p_2+1}^{p} t_j(\mathbf{h}_{i_1 i_2}) t_j(\mathbf{h}_{i_2 i_3}) t_j^*(\mathbf{h}_{i_1 i_3})$$

and

$$t_j(\mathbf{h}_{i_1 i_2}) = u_j(\mathbf{h}_{i_1 i_2}) \exp[i\beta_j(\mathbf{h}_{i_1 i_2})].$$

Note that all the terms of the form:

$$\langle t_j(\mathbf{h}_{i_1 i_2}) t_j(\mathbf{h}_{i_2 i_3}) t_j^*(\mathbf{h}_{i_1 i_3})\rangle_{\text{orient.}}$$

are calculated using the $B(z,t)$ formula (Hauptman, 1965), in a similar way to that described by Main (1976).

(*b*) *Space group* $P\bar{1}$

Considering terms only up to triplet order in a power series of Bessel functions and including only the 'primary' triplets, that is those formed by two generators, we finally obtained:

$$\prod_{j=1}^{p/2} q_j(\rho_{01}, \ldots, \rho_{i_1 i_2}, \ldots)$$

$$\simeq \exp\Big[-\tfrac{1}{2}\sum^{m\,\prime\prime\prime} \rho_{i_1 i_2}^2 - 2i \sum^{m\,\text{iv}} (Q_{0i_2, i_2 i_3, 0i_3} \rho_{0i_2} \rho_{i_2 i_3} \rho_{0i_3}$$

$$\times \cos q_{0i_2, i_2 i_3, 0i_3} + Q_{0i_2, i_3 i_2, 0i_3} \rho_{0i_2} \rho_{i_3 i_2} \rho_{0i_3} \cos q_{0i_2, i_3 i_2, 0i_3})\Big].$$

It is very important to note that in $P\bar{1}$:

$$\sum_{j=1}^{p_1/2} u_j^2(\mathbf{h}) + \sum_{j=p_1/2+1}^{p_2/2} \langle u_j^2(\mathbf{h})\rangle_{\text{orient.}} + \sum_{j=p_2/2+1}^{p/2} u_j^2(\mathbf{h}) = \tfrac{1}{2}.$$

The expressions for $Q_{0i_2, i_2 i_3, 0i_3} \exp(iq_{0i_2, i_2 i_3, 0i_3})$ and $Q_{0i_2, i_3 i_2, 0i_3} \exp(iq_{0i_2, i_3 i_2, 0i_3})$ are given by:

$$Q_{0i_2, i_2 i_3, 0i_3} \exp(iq_{0i_2, i_2 i_3, 0i_3})$$

$$= \sum_{j=1}^{p_1/2} t_j(\mathbf{h}_{0i_2}) t_j(\mathbf{h}_{i_2 i_3}) t_j^*(\mathbf{h}_{0i_3})$$

$$+ \sum_{j=p_1/2+1}^{p_2/2} \langle t_j(\mathbf{h}_{0i_2})\, t_j(\mathbf{h}_{i_2 i_3}) t_j^*(\mathbf{h}_{0i_3})\rangle_{\text{orient.}}$$

$$+ \sum_{j=p_2/2+1}^{p/2} t_j(\mathbf{h}_{0i_2}) t_j(\mathbf{h}_{i_2 i_3}) t_j^*(\mathbf{h}_{0i_3})$$

and

$$Q_{0i_2, i_3 i_2, 0i_3} \exp(iq_{0i_2, i_3 i_2, 0i_3})$$

$$= \sum_{j=1}^{p_1/2} t_j(\mathbf{h}_{0i_2}) t_j^*(\mathbf{h}_{i_3 i_2}) t_j(\mathbf{h}_{0i_3})$$

$$+ \sum_{j=p_1/2+1}^{p_2/2} \langle t_j(\mathbf{h}_{0i_2}) t_j^*(\mathbf{h}_{i_3 i_2}) t_j(\mathbf{h}_{0i_3})\rangle_{\text{orient.}}$$

$$+ \sum_{j=p_2/2+1}^{p/2} t_j(\mathbf{h}_{0i_2}) t_j^*(\mathbf{h}_{i_3 i_2}) t_j(\mathbf{h}_{0i_3})$$

and

$$t_j(\mathbf{h}_{i_1 i_2}) = u_j(\mathbf{h}_{i_1 i_2}) \exp[i\beta_j(\mathbf{h}_{i_1 i_2})].$$

APPENDIX II

The correlation between the observed and predicted densities can be approximated (Podjarny *et al.*, 1976) by the correlation between predicted and observed structure factors, $C(E_{\mathbf{h}}^p, E_{\mathbf{h}})$.

For small values of $(\varphi_{\mathbf{h}}^p - \varphi_{\mathbf{h}})$ this correlation is related to RMS4 by

$$C(E_{\mathbf{h}}^p, E_{\mathbf{h}}) = C(|E_{\mathbf{h}}^p|, |E_{\mathbf{h}}|)(1 - \text{RMS4}/2).$$

References

CASTELLANO, E. E., PODJARNY, A. D. & NAVAZA, J. (1973). *Acta Cryst.* A**29**, 609–615.

FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* A**34**, 517–525.

HAUPTMAN, H. (1965). *Z. Kristallogr.* **121**, 1–8.

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

HEINERMAN, J. J. L., KRABBENDAM, H. & KROON, J. (1979). *Acta Cryst.* A**35**, 101–105.

KARLE, J. & HAUPTMAN, H. (1956). *Acta Cryst.* **9**, 635–651.

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

MAUGUEN, Y. (1979). Thèse. Univ. de Paris VI.

PODJARNY, A. D., SCHEVITZ, R. W. & SIGLER, P. B. (1981). *Acta Cryst.* A**37**, 662–668.

PODJARNY, A. D. & YONATH, A. (1977). *Acta Cryst.* A**33**, 655–661.

PODJARNY, A. D., YONATH, A. & TRAUB, W. (1976). *Acta Cryst.* A**32**, 281–292.

RANGO, C. DE, MAUGUEN, Y. & TSOUCARIS, G. (1975). *Acta Cryst.* A**31**, S21.

SAYRE, D. (1974). *Acta Cryst.* A**30**, 180–184.

SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S. H. (1977). *Acta Cryst.* A**33**, 800–804.

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