Acta Cryst. (1976). A**32**, 281

Application of Multivariate Distribution Theory to Phase Extension for a Crystalline Protein

By A. D. Podjarny, A. Yonath and W. Traub

Department of Structural Chemistry, Weizmann Institute of Science, Rehovot, Israel

(*Received* 29 *May* 1975; *accepted* 4 *October* 1975)

As a test of the usefulness of matrix methods in extending phase information of proteins, the phases of triclinic lysozyme were calculated in the range from 3·3 to 2·5 Å (2000 phases) using as a starting set the phases out to 3·3 Å (1500 phases) and all the amplitudes. Observed ($|F_{\rm obs}|$) and calculated ($|F_{\rm modl}|$) amplitudes were tested separately. The agreement of the predicted phases and the original ones was studied, and was observed to depend strongly on the normalized value of the corresponding structure factor, $|E|$. Different methods for assessing the quality of the predicted phases were considered, and the correlation between electron density maps corresponding to predicted and original phases was selected as the most indicative of usefulness in structure determination. Predictions with the two different starting sets were studied using this correlation factor. A correlation of 0·5 between the predicted and original maps was obtained for the $F_{\rm modl}$, $\alpha_{\rm modl}$ case, and a correlation of 0·35 for the $F_{\rm obs}$, $\alpha_{\rm modl}$ case. Both correlations are significant.

1. Introduction

Direct methods are now a well established way of solving the phase problem in determinations of crystal structures with less than about 100 atoms in the asymmetric unit. In particular, the tangent formula (Karle & Karle, 1966) has proved to be very useful both for phase determination *ab initio* and for extending phase information from an initial set of known phases. However, applications of this method to protein structures (Reeke & Lipscomb, 1969; Weinzierl, Eisenberg & Dickerson, 1969; Coulter & Dewar, 1971; Destro, 1972) have met with only limited success. Perhaps the most successful example so far reported is the case of carp myogen for which phase refinement and extension from 2·0 Å to 1·85 Å resolution using the tangent formula led to some improvement in the electron density map (Hendrickson & Karle, 1973). Sayre (1974), using Sayre's formula (Sayre, 1953), achieved a far more striking improvement in the electron density map of rubredoxin by phase extension from 2·5 Å to 1·5 Å, but this procedure apparently requires near atomic resolution to begin with and proved unsuccessful when applied to a starting set of 3·0 Å resolution phases.

We have been concerned with developing an effective method for extending phase information to high resolution from a set of known phases out to 3 Å or 4 Å. This would be appropriate to the quite common situation in protein crystallography, where X-ray data for native protein crystals may extend to quite high resolution, but the standard method of multiple isomorphous replacement gives reliable phases only for relatively low-order reflexions because of imperfect isomorphism. In this work we have made use of covariance matrices (Tsoucaris, 1970*a*) which give expression to the relationship between structure factors in molecular crystals. In principle such a matrix should have an order greater or equal to the number of atoms in the asymmetric unit. In practice such a high order is very difficult to attain, and the limit is set by the number of available structure factors. Since Castellano, Podjarny & Navaza (1973) have shown that it is possible to set a fraction of the matrix elements equal to zero, the method can be used in practice for phase extension. The tangent formula is in effect a covariance matrix of order three (Tsoucaris, 1970*b*), and therefore less powerful than full matrices in predicting phases.

The phase predictions are made in terms of a set of generating reflexions, chosen according to criteria which are discussed below, and treated as random variables. The phase of any one of these reflexions can be determined from its statistical regression on all the other reflexions of the generating set. For this the E's of the generating reflexions are intercorrelated to generate a covariance matrix (U).

This Hermitian matrix is obtained from the complex E's phased by isomorphous replacement or by structure factor calculations, according to $U_{ij}=\langle {\rm E}(\mathbf{h}_j,\mathbf{r})\,.\,E^*(\mathbf{h}_k,\mathbf{r})\rangle_{\mathbf{r}}={\rm U}(\mathbf{h}_j-\mathbf{h}_k)$† (in space group $P1$) for indices $\mathbf{h}_j$ and $\mathbf{h}_k$ in the generating set, the average being taken over the atomic coordinates $\mathbf{r}$ which are considered random variables (Castellano *et al.*, 1973). Filling the $\mathbf{U}$ matrix implies *a priori* knowledge of $N(N-1)/2$ unitary structure factors. This matrix is then inverted to obtain the inverse matrix (**D**). Appropriate elements of the matrix and the E's of the generating reflexions are then used to calculate a unimodal one-dimensional probability distribution for the unknown phase according to formula (1) of §3. If one wishes to refine imprecisely known phases, one may include these reflexions to constitute the bulk or even all of the

† This equation is equivalent in $P1$ to equation 4 (de Rango, Tsoucaris & Zelwer, 1974), but derived differently.

generating set, but for phase extension one E for each unknown phase is added in turn to the generating reflexions.

We have tested the usefulness of this method of phase extension on triclinic lysozyme, the structure of which has recently been determined in our laboratory at 2·5 Å resolution with $R=35\%$ (Moult *et al.*, 1976). We used a set of observed structure amplitudes $F_{\rm obs}$ and phases calculated from the final lysozyme structure $\alpha_{\rm modl}$ to predict phases which were assumed unknown. We also made similar predictions starting with a set of $F_{\rm modl}$ and $\alpha_{\rm modl}$ in order to gauge the influence of experimental errors. The accuracy of the predictions was estimated in terms of the mean standard deviation between the values predicted for a set of phases and the corresponding 'correct' values calculated from the atomic coordinates of the structure.

Our investigation proceeded through the following stages:

(1) Covariance matrices were constructed using several sets of $F_{\rm modl}$ and $\alpha_{\rm modl}$ corresponding to various matrix orders (*i.e.* numbers of generating reflexions), occupancies (*i.e.* proportions of off-diagonal matrix elements not equal to zero) and lower limits for the modulus (E) of the generating reflexions. Phases were calculated for all the generating reflexions and mean standard deviations from the $\alpha_{\rm modl}$'s determined for the various cases. This investigation served to indicate optimum values for these three parameters and hence appropriate 'selection rules' for the generating reflexions.

(2) A covariance matrix was constructed using $F_{\rm modl}$ and $\alpha_{\rm modl}$ values from reflexions with spacings greater than 3·3 Å only. This matrix was then used to predict the phases of all reflexions between 2·5 Å and 3·3 Å and their accuracy was determined by comparison with the corresponding $\alpha_{\rm modl}$ values. This range of resolution was chosen because it implies an appreciable number of newly determined phases, provides a covariance matrix of sufficient occupancy to ensure a solution, and falls in the medium resolution range of great practical interest.

(3) Phase extension was carried out as in stage (2), but using $F_{\rm obs}$ and $\alpha_{\rm modl}$ values for the generating set and $F_{\rm obs}$ values for reflexions whose phases were predicted.

(4) The efficacy of the phase extension was further tested in direct space. An electron density map was calculated out to 2·5 Å resolution using $F_{\rm modl}$ values, together with $\alpha_{\rm modl}$'s for reflexions out to 3·3 Å and phases predicted in stage (2) for reflexions between 2·5 Å and 3·3 Å. This map was compared with 3·3 Å and 2·5 Å resolution maps, both prepared using $F_{\rm modl}$ and $\alpha_{\rm modl}$.

(5) A comparison similar to that outlined in (4) was carried out using $F_{\rm obs}$ instead of $F_{\rm modl}$, and the phases predicted in (3). The map incorporating the phase extension, which closely approximates to a practical situation, was found to show more correct detail than the 3·3 Å map calculated from $F_{\rm obs}$'s and $\alpha_{\rm modl}$'s. It was also found that there was a strong positive correlation in the electron densities contributed by the 2·5 Å to 3·3 Å reflexions using $F_{\rm obs}$ with $\alpha_{\rm modl}$ values and using $F_{\rm obs}$ values with phases predicted in stage (3).

2. Notation

$\sigma(x)$: Mean standard error of the variable x.

α: Phase of a structure factor.

$\alpha_N(i)$: Phase angle variable with a probability distribution, the mean of which is taken as the predicted phase for $E_N(i),\alpha_{\rm pred}$. This mean is also referred to as α_A, as it is the phase of $A_N(i)$.

$\alpha_{\rm modl}(i)$: Phase of $E_N(i)$, calculated from atomic coordinates.

F: Modulus of a structure factor: Either $F_{\rm obs}$ (experimentally measured) or $F_{\rm modl}$ (calculated from an atomic model).

$|E|$: Modulus of a statistically normalized structure factor.

E_j,E_k: Set of generating structure factors, normalized. Number$=N-1$. Known phase.

R_g: Reciprocal of smallest spacing represented by generating reflexions of known phase, Å^{-1}.

$E_N(i)$: Set of structure factors (normalized) whose phases are to be predicted. $i=1\ldots M$, M being the number of phases to be predicted.

R_p: Reciprocal of smallest spacing represented by reflexions whose phase is to be predicted, Å^{-1}.

$\mathbf{U}_{jk}$: Covariance of normalized structure factors E_j and E_k=element of submatrix $\mathbf{U}_{N-1}$ $j,k=1\ldots N-1$.

$\mathbf{U}^*_{Nj}=\mathbf{U}_{jN}(i)$: Element of last column of covariance matrix $\mathbf{U}$, corresponding to prediction of structure factor $E_N(i)$.

$\mathbf{U}_{jN}(i)$: is calculated as $\mathbf{U}(\mathbf{h}_j-\mathbf{h}_N)$, where $\mathbf{h}_j$ and $\mathbf{h}_N$ are the Miller indices of $\mathbf{E}_j$ and $\mathbf{E}_N$, respectively. [See Castellano *et al.*, (1973) equation (2.4), and apply for space group $P1$. $\mathbf{U}$ is denoted there as σ.]

R_k: Reciprocal of smallest spacing represented by reflexions whose phase is known *a priori*, Å^{-1}.

T: Total number of reflexions whose phase is known *a priori*.

$\mathbf{U}(i)$: Covariance matrix of all generating reflexions, including $E_N(i)$ as the last element. Its submatrix $\mathbf{U}_{N-1}$ remains constant as i changes.

$\mathbf{D}(i)$: Inverse of $\mathbf{U}(i)$. Contains submatrix $\mathbf{D}_{N-1}$ and last column and row $\mathbf{D}_{jN}$ and $\mathbf{D}_{Nj}$.

I_0,I_1: Zero and first-order modified Bessel functions.

P: Probability distribution function.

K_1, K_2: Normalization constants.

$A_N(i)$: $-\sum_{k-1}^{N-1} 2\mathbf{D}_{Nk}(i)E_k$ = predicted structure factor.

$B(i)$: $|A_N(i)| \,.\, |E_N(i)|$.

$w(i)$: $\cos\,[\alpha_A(i)-\alpha_N(i)]$.

$\bar{w}(i)$: Expected value of $w(i)$

$\bar{\bar{w}}$: Mean value of $\bar{w}(i)$, averaged over i.

$\mathbf{h}_j, \mathbf{h}_k$: Miller indices of normalized structure factors E_j and E_k.

rt: R_g/R_k.

C: Correlation between electron densities.

MI: Mean correlation. Ideal (w) weighting.

MR: Mean correlation. No weighting.

MW: Mean correlation. $\bar{w}$ weighting.

G: Number of electron density map grid points.

ϱ: $\varrho(x,y,z)$: Electron density.

Note that index i is sometimes not printed, but still implied.

3. Theory of errors in phase determination

The errors in phase determination depend on the set of generating reflexions, in so far as these represent the whole input of the method and determine its progress. It has been shown that the multidimensional probability distribution for a set of normalized structure factors is given by $P(E_1\ldots E_j,E_k\ldots E_N)=K_1 \,.\, \exp - \sum_{jk=1}^{N} E_j^* D_{jk} E_k$ [de Rango *et al.*, 1974, equation (5*b*)] provided the central limit theorem holds (Tsoucaris, 1970*a*; Castellano *et al.*, 1973). Now, the probability of the Nth generating reflexion having phase α_N,† becomes after normalization and collection of all the terms containing α_N

$$P(\alpha_N)=P(\alpha_N|E_1,\ldots E_j,E_N-1,|E_N|,\mathbf{U})=\frac{1}{2\pi I_0(B)} \times \exp\,[B\cos\,(\alpha_N-\alpha_A)] \quad (1)‡$$

where

$$A_N=|A_N|\,\exp\,(i\alpha)\; A=\sum_{j=1}^{N-1} -2\mathbf{D}_{Nj}E_j$$

and

$$B=|E_N|\,.\,|A_N|$$

[de Rango *et al.*, 1974, equation (10)].

This last formula represents a unimodal, one-dimensional probability distribution, whose shape is shown in Fig. 1. It has a maximum for $\alpha_N=\alpha_A$ which for this unimodal distribution represents both the centroid and the most probable phase, and the dispersion is a function of B, the sharpness increasing with B. The mean standard deviation of α_N is defined by

$$\langle\sigma(\alpha_N)\rangle^2=[\tfrac{1}{2}\pi I_0(B)]\int_{\alpha_A-\pi}^{\alpha_A+\pi}(\alpha_N-\alpha_A)^2 \times \exp\,[B\cos\,(\alpha_N-\alpha_A)]\mathrm{d}\alpha_N\,. \quad (2)$$

For an equally precise and more simple treatment, it is convenient to calculate $\bar{w}=\text{Mean}\,[\cos\,(\alpha_N-\alpha_A)]$ as a measure of dispersion, as

$$\bar{w}=\langle\cos\,(\alpha_N-\alpha_A)\rangle=[\tfrac{1}{2}\pi I_0(B)]\int_{\alpha_A-\pi}^{\alpha_A+\pi}\cos\,(\alpha_N-\alpha_A) \times \exp\,[B\cos\,(\alpha_N-\alpha_A)]\mathrm{d}\alpha_N=I_1(B)/I_0(B) \quad (3)$$

(McLachlan, 1955).

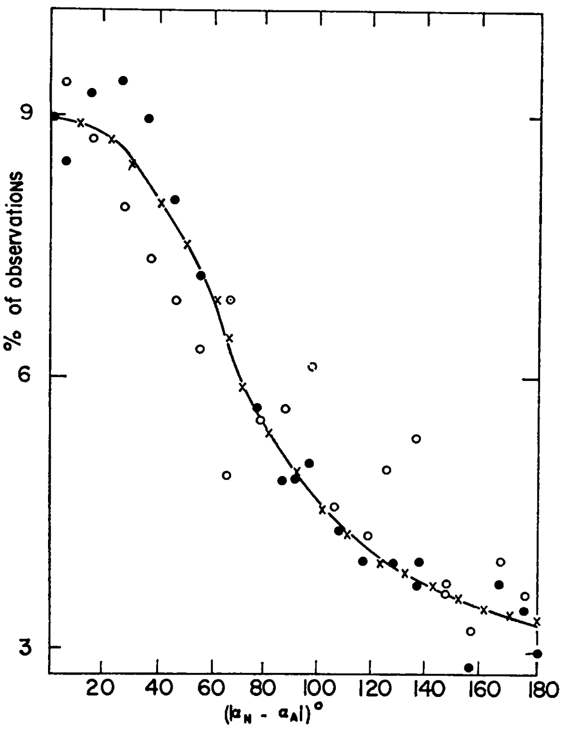

Fig. 1. Distribution of errors in the phase extension of the B range from 0 to 2 in both F_{modl} and F_{obs} cases, for triclinic lysozyme. Theoretical curve was calculated from the equation $P(\alpha_N-\alpha_A)=[2\pi I_0(B)]^{-1}\exp\,[B\cos\,(\alpha_N-\alpha_A)]$ for $B=0{\cdot}5$ ● = Exact data; ○ = observed data; × = theoretical curve.

† This phase, which could be any kth phase, is arbitrarily taken to be the last in order to simplify later discussion.

‡ To derive this formula, we start from

$$P(E_1\cdots E_N)=K_1\times\exp-\Big(\sum_{jk=1}^{N-1}E_j^*\mathbf{D}_{jk}E_k+\sum_{j=1}^{N-1}E_j^*\mathbf{D}_{jN}E_N+\sum_{k=1}^{N-1}E_N^*\mathbf{D}_{Nk}E_k+|E_N|^2\mathbf{D}_{NN}\Big),$$

isolate the terms dependent on α_N

$$\begin{aligned}P(\alpha_N)&=K_2\times\exp-[\text{real}\,(E_N^{*2}\sum_{k=1}^{N-1}\mathbf{D}_{Nk}E_k)]\\&=K_2\times\exp\,[|E_N|\,|A_N|\cos\,(\alpha_N-\alpha_A)]\\&=K_2\times\exp\,[B\cos\,(\alpha_N-\alpha_A)],\end{aligned}$$

then normalize

$$\begin{aligned}1=\int_0^{2\pi}P(\alpha_N)\mathrm{d}\alpha_N&=K_2\int_0^{2\pi}\exp\,[B\cos\,(\alpha_N-\alpha_A)]\mathrm{d}\alpha_N\\&=K_2\times2\pi I_0(B)\end{aligned}$$

from which $K_2=1/2\pi I_0(B)$ and (1) follows.

A similar measure of dispersion, the figure of merit m, has been used in Blow & Crick's (1959) treatment of multiple isomorphous replacement, and has proved very useful. We shall refer to $\bar{w}$ as the weight of a given phase prediction, as we shall later use it as a weighting factor to improve our maps in much the same way as the figure of merit, m, is used.

The limits for $\bar{w}$ are:

When $B \rightarrow \infty$, $\bar{w} \rightarrow 1$, corresponding to $\alpha_A = \alpha_{\text{modl}}$ (complete prediction) when $B \rightarrow 0$, $\bar{w} \rightarrow 0$, corresponding to $\alpha_A - \alpha_{\text{modl}} = \pm\pi/2$ (no prediction).

The relationship between B and the phase dispersion is defined by both (2) and (3). Formula (3) is more useful in practice, as (2) involves integral calculus.

It is useful in order to understand the character of B to approximate formula (1) for high values of B. This sharpens the distribution in it, rendering meaningful only the zone of integration where $(\alpha_N - \alpha_A)$ is small. It is therefore valid to use the approximation $\cos(\alpha_N - \alpha_A) = 1 - (\alpha_N - \alpha_A)^2/2$ which transforms the probability distribution of α_N into a Gaussian, $K_2 \exp[-B(\alpha_N - \alpha_A)^2/2]$. The variance of this Gaussian is $\sigma = B^{-1/2}$, or $B = \sigma^{-2}$. Thus, B is simply the inverse of the mean square error of the phase prediction, squared. This approximation implies an error of about $(\alpha_N - \alpha_A)^4/24$, which means that it is possible to use it as a comparison within the range of error of 0 to 1 radian (0° to 57°) with error of less than $\frac{1}{24}$ (4%) in the approximation of $\cos(\alpha_N - \alpha_A)$. Within the same range and with a similar error, we can equate:

$$\langle\cos(\alpha_N - \alpha_A)\rangle = \langle 1 - (\alpha_N - \alpha_A)^2/2\rangle = 1 - \langle(\alpha_N - \alpha_A)^2\rangle/2$$

which implies

$$\bar{w} = 1 - \sigma^2/2 = 1 - (1/2B) = (2B - 1)/2B \qquad (3a)$$

which, as will be shown later, is a good approximation to formula (3). Thus, it is clear that minimizing the error $(\alpha_{\text{modl}} - \alpha_{\text{pred}})$ or maximizing $\bar{w}$ is equivalent to maximizing B. Now considering the whole set of predicted phases, the best set of conditions corresponds to the maximum mean value of $\bar{w}$, averaged over all the predictions. If we define the mean value of $\bar{w}$ as

$$\overline{\overline{w}} = \frac{\sum_{i=1}^{M} \bar{w}|E_N(i)|^2}{\sum_{i=1}^{M} |E_N(i)|^2} = \frac{\sum_{i=1}^{M} \langle\cos[\alpha_N(i) - \alpha_A(i)]\rangle |E_N(i)|^2}{\sum_{i=1}^{M} |E_N(i)|^2}$$

$$= \langle \sum_{i=1}^{M} \cos[\alpha_N(i) - \alpha_A(i)]|E_N(i)|^2 / \sum_{i=1}^{M} |E_N(i)|^2\rangle$$

[where the summation i extends over all the predicted reflexions, and $\bar{w} = \bar{w}(i)$ for reflexion (i)], we have to maximize this function to obtain the best set of conditions. This corresponds to maximizing the set $\bar{w}(i)$ and hence, to maximizing $B(i)$. The reason for this particular weighting scheme in averaging $\bar{w}$ will be clear when we consider the use of correlations between electron density maps in assessing the quality of predicted maps. We will then see that the real correlation, MR, defined by

$$\text{MR} = \frac{\sum_i \cos[\alpha_{\text{modl}}(i) - \alpha_A(i)]|E_N(i)|^2}{\sum_i |E_N(i)^2|} \qquad (4)$$

represents the correlation between the electron density map with exact phases and the electron density map with predicted phases. Thus $\bar{w}$ equals the expected value of MR, which means that in terms of electron density maps maximizing the set $B(i)$ maximizes the correlation between predicted and correct electron density maps.

To perform this maximization we have to maximize $B(i)$ for each reflexion. This, in turn, recalling that

$$B(i) = |E_N(i)|\ |\sum_{j=1}^{N-1} 2\mathbf{D}_{Nj}E_j|\ , \qquad (5)$$

implies maximizing $|E_j|$, $|E_N(i)|$, N and $\mathbf{D}_{Nj}$, as well as aligning the terms in the summation. However, if we assume that the phases in the summation are random, we can neglect the last condition.

To maximize E_j we shall choose generating reflexions with structure factors having large moduli. For measuring the maximization of the elements $\mathbf{D}_{Nj}$, we express them as (see §5):

$$\mathbf{D}^*_{Nj}(i) = -\sum_{k=1}^{N-1} (\mathbf{U}^{-1}_{N-1})_{jk}\mathbf{U}_{kN}(i)\mathbf{D}_{NN}(i)$$

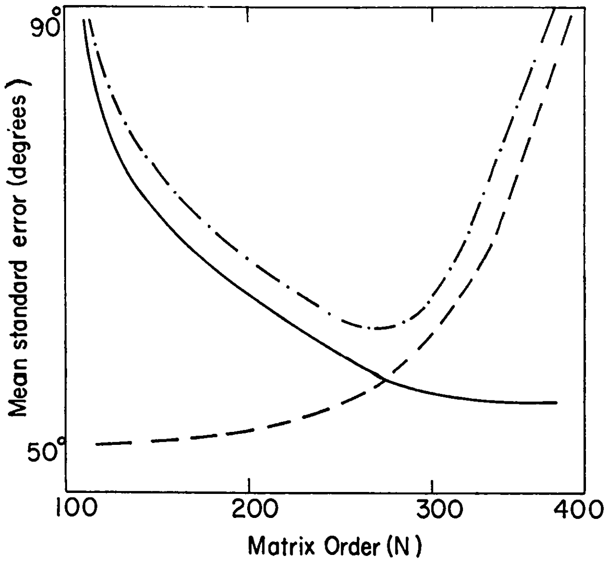

Fig. 2. Schematic representation of the influence on the final error in phase prediction of det $(\mathbf{U}_{ij})$ and delta det $(\mathbf{U}_{ij})$ as functions of the matrix order. Experimental results are shown in Fig. 4. ———— Error due to det $(\mathbf{U}_{ij})$. -------- Error due to delta det $(\mathbf{U}_{ij})$. –·–·– det $(\mathbf{U}_{ij})$ & delta det $(\mathbf{U}_{ij})$. Total error.

which enables us to put the summation in (5) as a quotient of determinants whose denominator is det $[\mathbf{U}(N-1)]$. Therefore, we see that maximizing $\mathbf{D}_{Nj}$ implies taking det $[\mathbf{U}(N-1)]$ close to zero. To do this, we may increase N, using the fact that a Karle–Hauptman determinant decreases with its order, and reaches zero when its order N equals the number of atoms in the unit cell (Karle & Hauptman, 1950). $\mathbf{U}_{ij}$ is not a Karle–Hauptman matrix, except for space group $P1$, but a Goedkoop matrix. Therefore, its determinant is zero when its order equals the number of atoms in the asymmetric unit (Goedkoop, 1950). In that way, maximizing $\mathbf{D}_{Nj}$ implies increasing the matrix order until it reaches the number of atoms in the asymmetric unit.

However, increasing N and decreasing det $(\mathbf{U}_{N-1})$ also increases the error introduced by the inversion of the matrix. Again taking det $(\mathbf{U}_{N-1})$ as a representative scalar, we see that

$$\Delta \det (\mathbf{U}_{N-1}^{-1}) = \frac{\Delta[\det (\mathbf{U}_{N-1})]}{\det [\mathbf{U}_{N-1}]^2} .$$

This implies that the error in the matrix inversion increases more rapidly than the inverse itself, and its effect is to decrease the $\overline{\overline{w}}$ and increase the error of the output phases (see Fig. 2). Therefore there exists a value of N, below that for which the determinant is zero, for which $\overline{\overline{w}}$ is a maximum.

In considering the maximization of $\bar{w}$, we have not yet introduced the fact that $\mathbf{U}_{jk}$ need not be completely full, but may contain zeros. These may correspond to structure factors of unknown value and this situation is comparable to an electron density from a Fourier summation of an incomplete set of structure factors. The consequences of this situation are twofold. Firstly, as the electron density is not necessarily positive, the positive definite character of $\mathbf{U}_{jk}$ need not hold for all values of N up to the number of atoms in the asymmetric unit, as would be the case for a complete matrix. In fact, the positive definite character is lost at an earlier stage, as is described in §4. This limits the theory, in its present form, to the range where the matrix is positive definite and therefore a proper covariance matrix. Further possibilities of extension, like filtering of negative eigenvalues, are under study. Secondly, another empirical parameter, which should be empirically optimized, is introduced. This is the occupancy of the matrix, defined as the proportion of non-zero elements. This is also determined by the generating reflexions, and we found experimentally that, if all the other parameters are kept constant, $\overline{\overline{w}}$ increases with the occupancy.

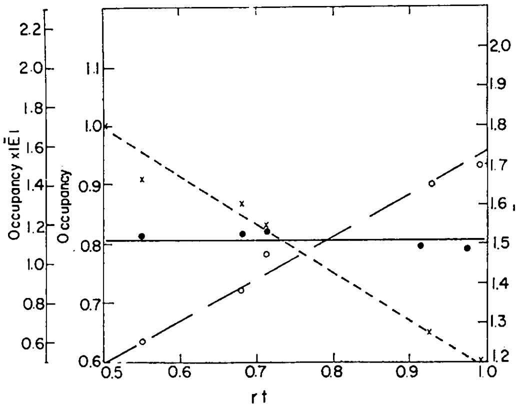

Fig. 3. Variation of occupancy and mean value of mod (E) with rt, as observed for matrices of order 400 built from 3Å resolution data. Note that the product of the two variables is approximately constant for the whole rt range. × ------- Observed occupancy of the matrix. ○ —— —— Observed mean values of the moduli of the generated reflexions. ● ———— Product of observed occupancy and mean moduli.

4. Experimental optimization

Thus, there are three parameters that influence $\overline{\overline{w}}$; the mean moduli of the E_j's, matrix occupancy and matrix order. The matrix order simply equals the number of generating reflexions. The occupancy, however, is related to the generating reflexions in a more complicated way, which requires some elaboration. The covariance matrix, in space group $P1$, is $\mathbf{U}_{jk}=\mathbf{U}(\mathbf{h}_j-\mathbf{h}_k)$, where $\mathbf{h}_j$ and $\mathbf{h}_k$ correspond to generating reflexions. Assuming, as is generally the case, that the whole set of structure factors with known phases lies within a known sphere in reciprocal space and that we choose the generating reflexions from within another sphere, the generating sphere, we conclude that the condition for 100% occupancy is that, for all $j,k,\mathbf{h}_j-\mathbf{h}_k$ must lie within the known sphere. This implies that the maximum modulus of $(\mathbf{h}_j-\mathbf{h}_k)$ is less or equal to the radius of the known sphere, that is to say

$$\max |\mathbf{h}_j-\mathbf{h}_k| \leq R_k = \text{radius of the known sphere} . \quad (6)$$

From the triangle law it follows that

$$\max |\mathbf{h}_j-\mathbf{h}_k| \leq \max (\mathbf{h}_j) + \max (\mathbf{h}_k) = 2\, R_g , \quad (7)$$

where R_g = radius of generating sphere. From (6), $\max |\mathbf{h}_j-\mathbf{h}_k| = R_k$ is a sufficient condition for 100% occupancy, so it follows that, if $2R_g = R_k$, (7) implies (6), and we have 100% occupancy. This reasoning can be immediately extended to space groups other than $P1$, using the fact that all matrices corresponding to symmetry transformations in reciprocal space are unitary, and replacing $\mathbf{h}_k$ by the symmetry-related reciprocal vector.

Defining $rt = R_g/R_k$ as a normalized variable for the radius of the generating sphere, we started from $rt = 0{\cdot}5$, which implies 100% occupancy. Increasing rt, we found that $rt = 1{\cdot}0$ implies 60% occupancy.

We have found that the variation of occupancy with rt is practically linear for rt between 0·5 and 1·0 (see Fig. 3).

The relationship between rt and the mean value of the moduli of the generating reflexions can be examined in a similar way. Assuming the values of mod (E) follow a normal Gaussian distribution, the percentage of reflexions in the generating sphere with mod (E) above some minimum value γ is a function of rt, the order of the matrix $N-1$, and the total number of reflexions in the known sphere T. To find this relationship, let us assume that the $N-1$ reflexions of greatest mod (E) within the generating sphere are taken as generating reflexions. Then the percentage (PG) that these constitute of the total number of reflexions in the generating sphere is:

$$\mathrm{PG}=N\,.\,100/K_3\,.\,\tfrac{4}{3}\pi R_g^3 \qquad (8)$$

where K_3 is a constant relating the number of reflexions to the volume in reciprocal space. In the known sphere, K_3 is defined by

$$T=K_3\,.\,\tfrac{4}{3}\pi R_k^3\,. \qquad (9)$$

Combining (8) and (9), it follows that

$$\mathrm{PG}=100\,.\,\frac{N}{T}\,.\,rt^{-3}\,.$$

We now use the fact that mod (E) is normally distributed to obtain the following formula that relates PG to mod (E), and hence mod (E) to rt, as:

$$\mathrm{PG}=\frac{100}{\sqrt{2\pi}}\int_{\gamma}^{\infty}\exp\left(-E^2/2\right)\mathrm{d}E=\frac{100\mathrm{erfc}\,(\gamma)}{\sqrt{2\pi}}$$

(for erfc see Abramowitz & Stegun, 1965)

$$\gamma=\mathrm{erfc}^{-1}[(\mathrm{PG})\,.\,\sqrt{2\pi}/100]\,.$$

$$\mathrm{Mean}\ [\mathrm{mod}\ (E)]=\frac{1}{\sqrt{2\pi}}\int_{\gamma}^{\infty}E\exp\left(-E^2/2\right)\mathrm{d}E$$

$$=\frac{1}{\sqrt{2\pi}}\exp\left[-\mathrm{erfc}^{-1}\left(\frac{N}{T}\,.\,rt^{-3}\,.\,\sqrt{2\pi}\right)\right].$$

The relation between mod (E) and rt, for an actual case, is shown in Fig. 3. It can be seen from that figure that changing rt from 0·5 to 1·0 causes the occupancy and mod (E) to change in opposite directions, thereby cancelling out their influences on $\overline{\overline{w}}$.

To decide which rt value was better, we predicted the phases of the generating reflexions for $rt=0{\cdot}6$ and $rt=1{\cdot}0$, varying in each case the matrix order.

The results of the two phase predictions are shown in Fig. 4 whereas Table 1 shows a comparison between the two cases taking into account only the reflexions common to both so as to avoid any spurious differences due to different phases having been predicted. This analysis suggests that the case of $rt=1{\cdot}0$ and $N=320$ is the best. However, it should be remarked that only a limited range of parameters was tested, and this result might well depend on the special conditions of the lysozyme case.

Table 1. *Comparison of mean errors between predicted and original phases for* $rt=1{\cdot}0$ *and* $rt=0{\cdot}6$ ($N=336$), *on the basis of* 62 *common reflexions*

rt	Mean error (°)
0·6	45
1·0	29

It is important to note that all the optimization of parameters for phase extension described above was performed in reciprocal space and with only 10% of the total available reflexions. Consequently we did not extend the comparison of original and predicted phases to include electron density maps.

5. Theory of phase extension

The preceding investigation showed us that formula (1) could be used to predict the phases of the generating reflections α_N and indicated optimum parameters for performing such predictions.

We then used these parameters to predict an unknown set of phases corresponding to reflexions between two spherical surfaces in reciprocal space with radii R_k and R_p. This could correspond to a practical situation where multiple isomorphous replacement has

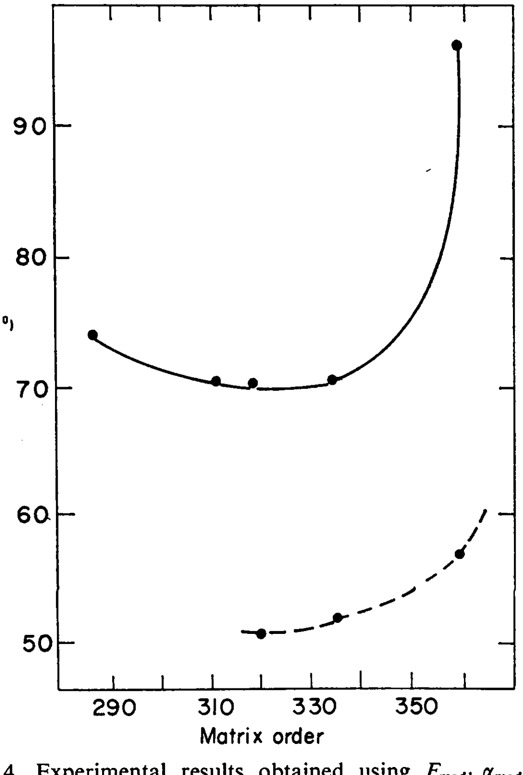

Fig. 4. Experimental results obtained using $F_{\mathrm{modl}},\alpha_{\mathrm{modl}}$ of triclinic lysozyme (out to 2·5 Å) that show the variation of mean standard error in the predicted phases of the generating reflexions as a function of matrix order for $rt=0{\cdot}6$ (————) and $rt=1{\cdot}0$ (— — —). The curve for the $rt=$ 0·6 case may be extended to take account of points not shown at 102° and 80° for matrix order 100 and 230 respectively.

yielded reliable phases out only to resolution R_k, whereas structure factors of the native protein are available out to resolution R_p.

In the phase extension procedure, the phase α_N must be the phase of a generating reflexion. However, whereas in the previous situation we included all the reflexions whose phases we predicted together among the generating reflexions, we cannot do so for phase extension, because of the large number of unknown phases (about 2000 in the case we treated), because these reflexions do not fulfil the optimum set of conditions, and because we need to know the phases of all but a few generating reflexions (all but one in our procedure) in order to calculate the unknown phases as a statistical regression upon the known ones. It should be noted that this calculation depends both on the covariance matrix and on the generating reflexions of known phase, and is therefore not equivalent to other procedures depending only on the covariance matrix, where the solution is the eigenvector corresponding to the largest eigenvalue of the matrix (Main, 1973). We may therefore consider two groups of generating reflexions. The large group of known phase fulfils the optimum set of conditions and remains constant for the whole phase extension procedure. The other consists of reflexions with phases to be predicted, taken a few at a time. The reflexions in the latter group will not, in general, have high mod (E), and their interactions with the rest will not necessarily have high occupancy. It is therefore desirable, in order to fulfil the optimum set of conditions, to reduce this second set of generating reflexions to the minimum, the ideal case being prediction for one reflexion at a time of phase $\alpha_N(i)$, where i is an index running through all the phases to be predicted.

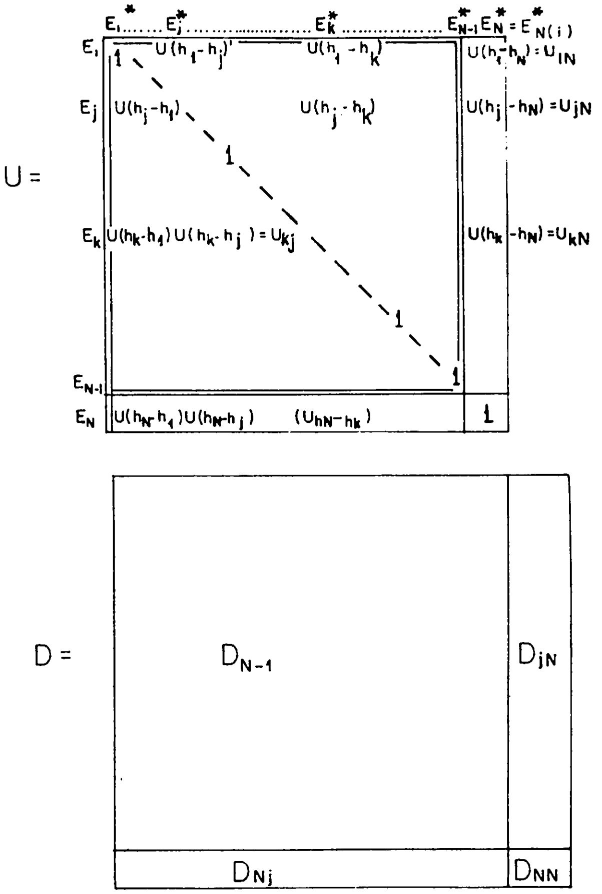

Fig. 5. Blocking of the matrices **U** and $\mathbf{D}=\mathbf{U}^{-1}$ for inversion purposes. Matrix **U** is generated in space group $P1$. The generating reflexions E_k are shown for clarity, but they do not form part of the covariance matrix **U**. The blocks of the matrix **D** are obtained from those in matrix **U** using formulae explained in the text. The part of matrix **U** which is framed by a double line is the $\mathbf{U}_{N-1}$ submatrix.

This is the procedure we used according to the scheme outlined in Fig. 5. An apparent difficulty is that it implies a different statistical problem (and therefore a different matrix inversion) for each phase to be predicted. However, all the different matrices (2000 in our case) differ only in their last row and column, and it is in fact possible to perform all the 2000 inversions in one step.

It can be seen from the basic formula

$$P[\alpha_n(i)] = \frac{1}{2\pi I_0(B)} \exp\,[|E_N(i)|\,.\,|A_N(i)|] \times \cos\,[\alpha_N(i) - \alpha_A\,(i)]$$

where

$$A_N(i) = -\sum_{j=1}^{N-1} \mathbf{D}_{Nj}(i) E_j$$

that all the information needed to predict $\alpha_N(i)$ is contained in $E_N(i)$ and in the various values of E_j and $\mathbf{D}_{Nj}(i)$ for $j=1$ to N. Though each of the 2000 phase predictions requires the calculation of different values of $\mathbf{D}_{Nj}(i)$, we can avoid having 2000 matrix inversions by considering the product of the matrix **U** and the inverse matrix **D** in terms of the blocks $N-1\times N-1$, $N\times 1$, $1\times N$ and 1×1 illustrated in Fig. 5. This product equals the unit matrix so that

$$\sum_{k=1}^{N-1} \mathbf{U}_{jk}\mathbf{D}_{ks} + \mathbf{U}_{jN}(i)\mathbf{D}_{Ns}(i) = \delta_{js} \quad j,s=1,\ N-1$$

$$\sum_{k=1}^{N-1} \mathbf{U}_{jk}\mathbf{D}_{kN}(i) + \mathbf{U}_{jN}(i)\ \mathbf{D}_{NN} = 0 \quad j=1,\ N-1$$

$$\sum_{k=1}^{N-1} \mathbf{U}_{Nk}(i)\mathbf{D}_{kN}(i) + \mathbf{U}_{NN}\mathbf{D}_{NN} = 1$$

which leads to the solution (Ayres, 1962)

$$\begin{aligned}
(\mathbf{D}^{(i)}_{N-1})_{jk} &= (\mathbf{U}^{-1}_{N-1})_{jk} \\
&\quad + \sum_{l,s} (\mathbf{U}^{-1}_{N-1})_{jl}\mathbf{U}_{lN}(i)\mathbf{D}^{(i)}_{NN}\mathbf{U}_{Ns}(i)\ (\mathbf{U}^{-1}_{N-1})_{sk} \\
\mathbf{D}^{(i)}_{NN} &= [\mathbf{U}^{(i)}_{NN} - \sum_{jk} \mathbf{U}_{Nj}(i)(\mathbf{U}^{-1}_{N-1})_{jk}\mathbf{U}_{kN}(i)]^{-1} \\
\mathbf{D}_{jN}(i) &= -\sum_{k} (\mathbf{U}^{-1}_{N-1})_{jk}\mathbf{U}_{kN}(i)\mathbf{D}^{(i)}_{NN} \\
\mathbf{D}^{(i)}_{Nj} &= \mathbf{D}^{*(i)}_{jN} \quad \text{and} \quad j,k,l,s=1,\ldots,N-1\,. \qquad (10)
\end{aligned}$$

The last line gives the values of $\mathbf{D}_{Nj}(i)$ required for phase prediction in terms of only one matrix inversion, $\mathbf{U}_{N-1}^{-1}$. $\mathbf{D}_{NN}$ is a scalar.

6. Practical applications of phase extension

It can be seen from formula (10) that the occupancy of the matrix column $\mathbf{U}_{jN}(i)$ directly affects the values of the inverse elements $\mathbf{D}_{jN}(i)$ and hence the size of mod (A) and the accuracy of the phase prediction $A_N(i)$ discussed above. This occupancy depends on the radius R_g of the sphere which encloses the generating reflexions E_j and the radii R_k and R_p which define a shell containing reflexions $E_N(i)$ whose phases we wish to predict. We chose R_k and R_p as the reciprocal of 3·3 Å and 2·5 Å respectively. This implies predicting 2000 phases starting from 1400 reflexions with known phases and gave an average occupancy for the matrix columns $\mathbf{U}_{jN}(i)$ of about 35%.

First we used $F_{\rm modl}, \alpha_{\rm modl}$ values and inverted the $\mathbf{U}_{N-1}$ matrix having only reflexions out to 3·3 Å and an occupancy of 59%. We then used the $\mathbf{D}$ matrix to predict the $N-1$ (in this case 336) generating reflexions. The mean standard error of these predicted phases is 65° (degrees).

We then added the various $\mathbf{U}_{jN}(i)$ columns, calculated values of $\mathbf{D}_{jN}(i)$ from equation (10) and hence derived $A_N(i)$ and predicted phases for the 2000 additional reflexions. An analysis of these extended phases for different B ranges is shown in Fig. 6 and a comparison of the observed error distribution with that expected from formula (1) is shown in Fig. 1. It is clear that the observed distribution agrees well with theoretical expectations.

We also compared observed $\bar{w}=$ Mean $[\cos(\alpha_{\rm modl}-\alpha_{\rm pred})]$ with the theoretical value calculated according to both the exact formula $\bar{w}=I_1(B)/I_0(B)$ and to the approximate formula $\bar{w}=1-(1/2B)$. The results are shown in Table 2, from which it can be seen that the expected and observed $\bar{w}$ values agree, and that the approximate formula is useful for $B>1{\cdot}5$.

Table 2. *Mean value of* $\cos(\alpha_{\rm modl}-\alpha_{\rm pred})$ *in different ranges of B*

$\alpha_{\rm pred}$ is the phase predicted in the shell 2·5 –3·3 Å from $F_{\rm modl}$ $\alpha_{\rm modl}$ data to 3·3 Å.

B range	Mean value of $\cos(\alpha_{\rm modl}-\alpha_{\rm pred})$	$I_1(B)/I_0(B)$	$1-(1/2B)$	Percentage of reflexions
0–2	0·26	0·42	0·50	86
2–4	0·61	0·83	0·84	12
4–6	0·77	0·90	0·90	1·4
6–8	0·78	0·94	0·94	0·3

These two tests show that the errors predicted by the theory agree with the observed errors for the $F_{\rm modl}$, $\alpha_{\rm modl}$ case. However, we also wished to define a mean weight for the predicted phases that would be roughly equivalent to the mean figure of merit used in the multiple isomorphous replacement method (Blow & Crick, 1959). This should be some mean of $\cos(\alpha_{\rm modl}-\alpha_{\rm pred})$ and we chose the mean square value of w, weighted with mod $[E(i)]^2$, as follows

$$({\rm MI})^2=\{\sum_{i=1}^{M} w^2(i)|E_N(i)|^2\}/\sum_{i=1}^{M}|E_N(i)|^2\,. \qquad (11)$$

The reason for this weighting is that it gives greatest importance to the reflexions that contribute most to the electron density, making MI a measure of the accuracy of the prediction in direct space, *i.e.* a correlation between the correct and predicted electron density maps. This correlation is defined as

$$\begin{aligned} C&=\langle(\varrho_1-\bar{\varrho}_1)\,.\,(\varrho_2-\bar{\varrho}_2)\rangle/\sigma(\varrho_1)\,.\,\sigma(\varrho_2)\\ &=\frac{\langle\varrho_1\varrho_2\rangle}{\sigma(\varrho_1)\sigma(\varrho_2)}-\frac{\bar{\varrho}_1\bar{\varrho}_2}{\sigma(\varrho_1)\sigma(\varrho_2)}\\ &=\left\{\frac{\sum\limits_i w(i)^2|E_N(i)|^2}{\sum\limits_i |E_N(i)|^2}\right\}^{\frac{1}{2}}-\frac{\bar{\varrho}_1\bar{\varrho}_2}{\sigma(\varrho_1)\sigma(\varrho_2)} \end{aligned} \qquad (12)$$

[where $\varrho_1(x,y,z)=$ electron density $=$ Fourier transform of $|E|\,.\exp(i\alpha_{\rm modl})$, $\varrho_2(x,y,z)=$ Fourier transform of

$$w|E|\exp(i\alpha_{\rm pred}),\ \langle\varrho\rangle=\frac{1}{G}\sum_{q=1}^{G}\varrho(xq,yq,zq)\,,$$

$G=$ number of grid points, and Fourier transforms are calculated from structure factors between 3·3 Å and 2·5 Å in (12)]. C is equal to MI provided that the second term is much smaller than the first, a condition that is generally met in practice.

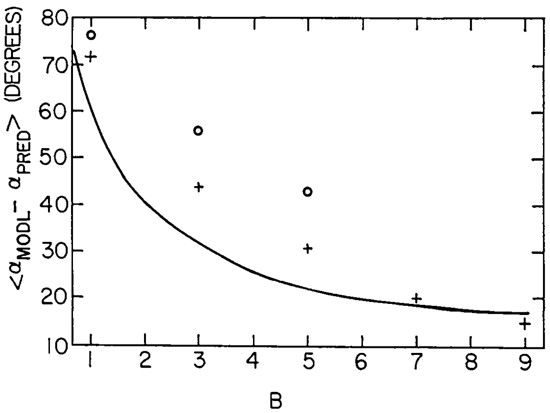

Fig. 6. Phase extension. Variation of $\langle\delta_\alpha\rangle=\langle(\alpha_{\rm modl}-\alpha_{\rm pred})\rangle$ for predicted phases (from 3·3 Å to 2·5 Å), matrix order = 336, as a function of B. $+=F_{\rm modl}$ data; $\circ=F_{\rm obs}$ data. The theoretical behaviour of $\langle\delta_\alpha\rangle$ for $B>1$, is represented by a full line

$$\langle\delta_\alpha\rangle=\frac{1}{\pi}\int_0^\pi \delta_\alpha\exp(B\cos\delta_\alpha){\rm d}\delta_\alpha\,.$$

This figure has similar format to Fig. 1 of de Rango, Mauguen &Tsoucaris (1975).

A calculation was made of this correlation factor between electron density maps calculated from data for the range 2·5 Å to 3·3 Å using $F_{\text{modl}}, \alpha_{\text{modl}}$ and $F_{\text{modl}}, \alpha_{\text{pred}}$ (see §7). This factor is independent of the scales and the zero-points of the maps, and measures the resemblance between the features of the two maps. As these maps are calculated using the extended phases from 3·3 to 2·5 Å, the value of MI gives a measure of the accuracy of the extended phases in terms of electron density. The value of MI for these predictions is 0·683. This calculation, for a map with 50000 grid points, is accurate to 1·3%, with a confidence level of 99·8%. When correlations are calculated between unrelated maps, they give zero to within 0·5%. For comparison, MI was calculated from formula (11) for tetragonal lysozyme structure factors with figures of merit derived from multiple isomorphous replacement (kindly supplied by Professor D. C. Phillips) and found to be 0·89.

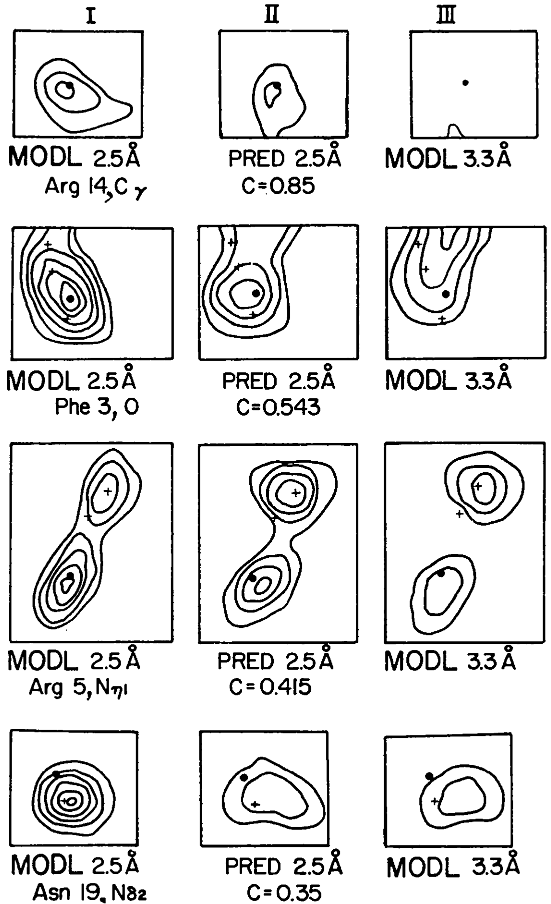

Fig. 7. Comparison of three electron density maps of triclinic lysozyme for several regions. The maps were constructed by using F_{obs} and (I) α_{modl} to 2·5 Å resolution, (II) α_{modl} to 3·3 Å resolution and α_{pred} for 2·5 Å to 3·3 Å resolution, (III) α_{modl} to 3·3 Å resolution. The correlation (C) was calculated for difference maps between (I)–(III) and (II)–(III) in the neighbourhood (sphere with 1·5 Å radius) of the atomic positions marked by (●). Contouring intervals for the 2·5 Å maps are 2 e Å^{-3} and for the 3·3 Å resolution map 1·5 e Å^{-3}. Atoms not included in the calculation of C are shown as (+).

The MI analysis is based on the individual weights $w = \cos(\alpha_{\text{modl}} - \alpha_{\text{pred}})$ for each structure factor. However, in a real case, the original phase α_{modl} is not known. Therefore we calculated the correlation between the predicted and original phases, without any weights, and this correlation was found to be MR = 0·44. It is also possible to weight the structure factors with predicted phases in calculating a map according to the formula $\bar{w} = I_1(B)/I_0(B)$ and this resulted in a correlation of MW = 0·48.

These correlation factors are essentially a measure of the degree of peak overlap in the two maps. Table 3 shows the value of the correlation between two Gaussian peaks, of width a, as a function of the peak-to-peak distance b. This is an idealized case, because the actual maps do not show resolution of individual atoms. In order to give an example of a real case, we took three electron density maps, all calculated from observed amplitudes (see §7), one corresponding to 3·3 Å resolution with model phases, another to 2·5 Å resolution with model phases, and the third to 2·5 Å resolution with model phases out to 3·3 Å and predicted phases from 3·3 Å to 2·5 Å, and chose several regions of electron density corresponding to different correlation values. These correlations were calculated only in the environments of particular atoms, the correspondence between the numerical values of C and the visual evidence of electron density at the atomic positions introduced by the predicted phases is illustrated in Fig. 7.

Table 3. *Theoretical correlation between two atoms that have Gaussian shape*

Width = a and distance between peaks = b. The theoretical value of correlations is exp $[0\cdot25\,(b/a)^2]$.

b/a	C
0·0	1·00
0·5	0·93
1·5	0·78
1·5	0·57
2·0	0·37
2·5	0·21
3·0	0·10
4·0	0·01

7. Test of the theory on experimental data and conclusions

All the above results were obtained using a set of model structure factors. As a further test of the theory we decided to use as input data observed structure factors for triclinic lysozyme and phases calculated from a model which gave an R value of 35% (Moult *et al.*, 1976).

The procedure followed is that described for the $F_{\text{modl}}, \alpha_{\text{modl}}$ case. We built and inverted a 336-order matrix, using only the structure factors out to 3·3 Å.

The occupancy was 57·5% with $rt = 1·0$, and the mean standard error for predicted phases of the generating reflexions was 75·1°.

Comparing this result with the $F_{\mathrm{modl}},\alpha_{\mathrm{modl}}$ case ($\sigma_\alpha = 65°$) we see that the error is larger, presumably because of observational errors inherent in the experimental data.

We then added $\mathrm{U}_{jN}(i)$ columns with average occupancy of 35% and calculated phase extensions. The results in terms of error distribution also appear in Fig. 5 which shows a comparison of errors for the two sets of phase extension. Fig. 1 compares the error distributions for various ($\alpha_{\mathrm{modl}} - \alpha_{\mathrm{pred}}$) ranges with the theoretical predictions.

We also compared corresponding electron density maps in the same way as for the exact case, but used only the MW correlation, which turned out to be 0·33 between $F_{\mathrm{obs}},\alpha_{\mathrm{pred}}$ and $F_{\mathrm{obs}},\alpha_{\mathrm{modl}}$ difference maps between 3·3 Å and 2·5 Å resolution. MW between the original $F_{\mathrm{obs}},\alpha_{\mathrm{modl}}$ map and the $F_{\mathrm{modl}},\alpha_{\mathrm{modl}}$ map is 0·89, so the mean weight of our predicted map as compared with the original one is $0·33/0·89 = 0·37$.

We also performed an analysis of correlations in the neighbourhoods of the various atoms. Analysis of the observed maps, calculated only with structure factors in the 3·3 to 2·5 Å range, showed that most of the extra density concentrated along side chains and the main chain carbonyls, as shown in Table 4.

This is in agreement with previous observations (North & Philips, 1969) that at 3·5 Å resolution polypeptide chains are observed as columns of continuous high density with prominent peaks marking the branching points of side chains at the alpha carbons of amino acid residues, whereas at higher resolution the most prominent peaks along the backbone are the peptide carbonyl groups. That is, alpha carbon positions are observed (and even overemphasized) at 3·5 Å, whereas carbonyls are not, because of the rounding effect of low resolution, and therefore the extra density added in going from 3·5 Å to 2·5 Å should diminish the alpha carbons and enhance the oxygen positions. Other easily recognizable features in a low-resolution map include the aromatic rings of tryptophan, tyrosine, phenylalanine and histidine. Methionine residues are also prominent, but they are not always clearly distinguishable from aromatic residues, whereas at higher resolution the high density peak of the sulphur is easily seen. Cystine disulphide bridges are also seen at low resolution, but often they cannot be distinguished from the main chain direction, which greatly complicates interpretation of the map (Kartha, 1967). Thus the main contributions to map interpretation of the high-resolution data include the location of carbonyl groups and hence the orientation of the peptide bond, the identification of disulphide bonds as such, and the identification of methionine residues, which are generally present in small numbers and are therefore very useful for following the amino-acid sequence within the map. Other extra features, like the flattening of aromatic rings and the appearance of lighter side chains, are also useful in the transition from the non-interpretable map to an interpretable one.

In the light of these considerations, we have analysed the correlations, for the different types of atoms, between the electron density contributed by the predicted phases and density derived from calculated phases, both with observed structure factors. Correlation for the main-chain atoms are shown as histograms

Table 4. *Mean value of density around each type of atom (out to a radius of* 1·5 Å) *contributed only by reflexions in the* 3·3 Å *to* 2·5 Å *shell to the* $F_{\mathrm{obs}},\alpha_{\mathrm{modl}}$ *map and corresponding correlations with the* $F_{\mathrm{obs}},\alpha_{\mathrm{pred}}$ *map*

Atom type	Total number of atoms	% of Atoms in aromatic rings*	Mean density	Mean correlation (MW)	Fraction with correlation > 0·5
C^α	129		−10·45	0·20	0·17
N	129		3·92	0·22	0·20
C	129		1·09	0·29	0·30
O	129		16·71	0·32	0·32
C^β	117		−1·63	0·29	0·21
C^γ	99		2·48	0·29	0·21
C^δ	71	50	0·22	0·23	0·26
C^ε	32	75	−3·94	0·30	0·25
C^ζ	29	66	−1·92	0·33	0·24
C^η	6	100	−0·87	0·30	0·33
N^δ	29		6·36	0·37	0·28
N^ε	25	35	0·01	0·26	0·36
N^ζ	6		4·40	0·44	0·50
N^η	22		0·21	0·33	0·36
O^γ	17		13·95	0·30	0·24
O^δ	14		11·54	0·38	0·28
O^ε	4		12·97	0·46	0·59
O^η	3		11·96	0·40	0·33
S^γ	8		11·81	0·46	0·50
S^δ	2		10·28	0·45	0·50

* Tryptophan, phenylalanine and tyrosine.

in Fig. 8, and corresponding data for all atoms are summarized in Table 4. This indicates that extra density is correctly located at the sulphur peaks, both for cysteine (S^{γ}) and for methionine (S^{δ}). Also much of the density is correctly located at main chain carbonyls, whereas low correlations at C^{α} positions simply indicate that these peaks were already established at 3·3 Å resolution. These features are among the most important for map interpretation, so that locating at least 32% of the oxygens and 50% of the sulphurs could be a significant aid in determining the backbone structure. Furthermore, side-chain atoms, especially oxygens and nitrogens at the ends of hydrophilic groups, which are not easily identifiable at 3·3 Å resolution also show up more clearly because of the additional electron density contributed by terms with predicted phases.

The aim of this work has been to find a solution to the phase problem using statistical procedures in what is perhaps the most difficult and critical range in structure determination of biological macromolecules. This so-called medium range covers from 4 Å to 2·5 Å and is often characterized by a failure of multiple isomorphous replacement to provide sufficiently good phases to start refinement based on an initial structural model. By applying the statistical approach to the test case of triclinic lysozyme, it has been shown that the method can provide useful structural information particularly in the regions of the protein side chains and the main-chain carbonyls which indicate the orientations of the peptide groups. Even though this information is not of the quality obtainable with good isomorphous phasing extending to 2·5 Å, it would surely have helped in map interpretation had multiple isomorphous replacement yielded good phases for lysozyme only out to 3·3 Å. Though, because of the special circumstances of this test case, we cannot conclude that this method would be generally applicable in other cases, it certainly does show a possible approach to the difficult problem of solving protein structures with limited isomorphous phases.

We are grateful to Drs Michael Levitt, John Moult and Paul Sigler for many helpful and interesting discussions and to Dr Anita Smilansky for her help in data processing. We would also like to express our appreciation of research grants from the Volkswagen Stiftung (Az 11790) and the Minerva Foundation.

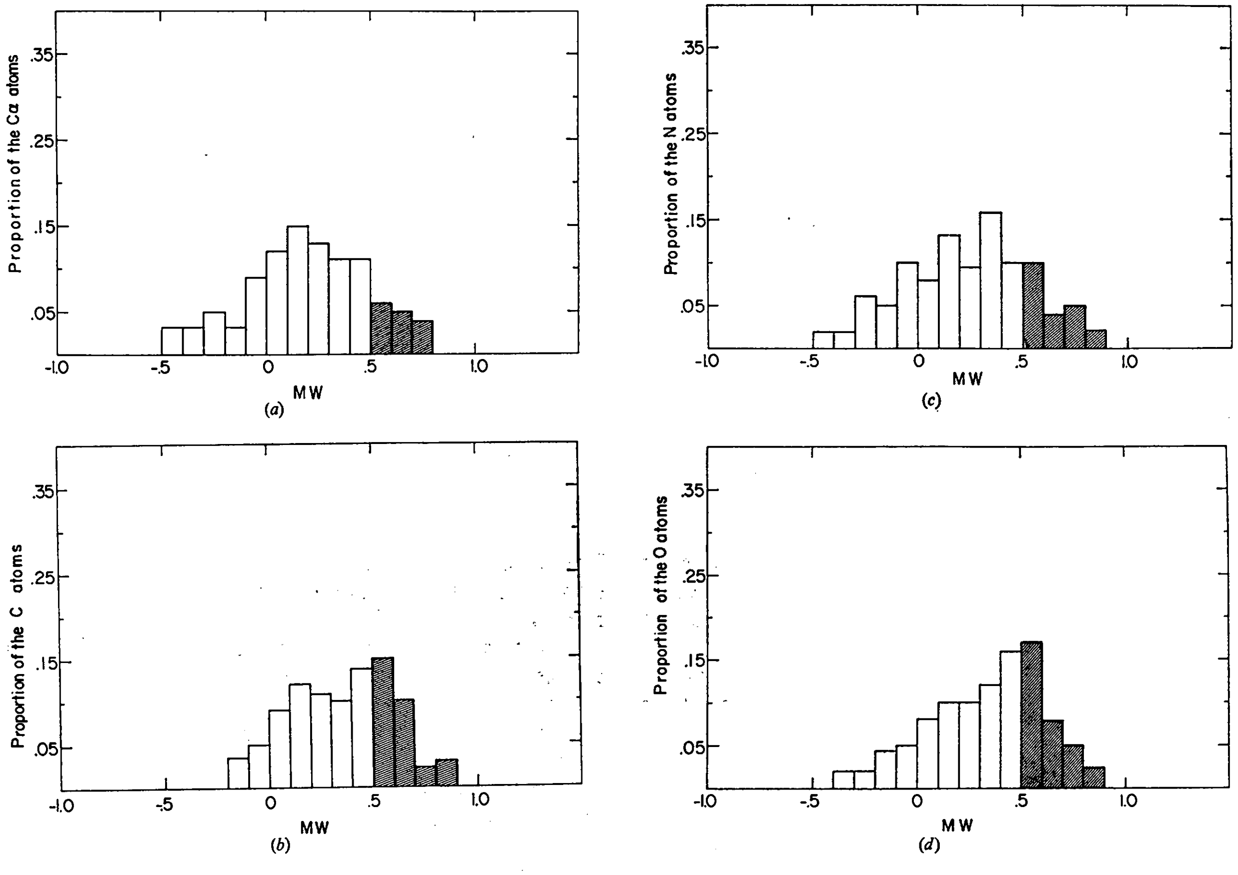

Fig. 8. Histograms of correlations (MW) for main-chain atoms between electron densities contributed by predicted and calculated phases for the F_{obs},α_{mod1} case. The proportion of total atoms having various correlations is shown, and the most significant correlations (MW > 0·5) are shaded.

References

ABRAMOWITZ, M., & STEGUN, I. A. (1965). *Handbook of Mathematical Functions*, p. 297. New York: Dover.
AYRES, F. (1962). *Matrices*, p. 58. Schaum Publ. Co.
BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794.
CASTELLANO, E., PODJARNY, A. & NAVAZA, J. (1973). *Acta Cryst.* A**29**, 609–615.
COULTER, C. L. & DEWAR, R. B. K. (1971). *Acta Cryst.* B**27**, 1730–1740.
DESTRÓ, R. (1972). *Report CECAM Workshop*, p. 9.
GOEDKOOP, J. A. (1950). *Acta Cryst.* **3**, 374–378.
HENDRICKSON, W. A. & KARLE, J. (1973). *J. Biol. Chem.* **243**, 3327–3340.
KARLE, J. & HAUPTMAN, H. (1950). *Acta Cryst.* **3**, 181–187.
KARLE, J. & KARLE, I. L. (1966). *Acta Cryst.* **21**, 849–859.
KARTHA, G. (1967). *Nature, Lond.* **214**, 234–330.
MCLACHLAN, N. W. (1955). *Bessel Functions*, p. 202. Oxford: Clarendon Press.
MAIN, P. (1973). *Commun. CECAM Symposium*, Holland. p. 14.
MOULT, J., YONATH, A., TRAUB, W., SMILANSKY, A., PODJARNY, A. D., SAYA, A. & RABINOVICH, D. (1976). *J. Mol. Biol.* In the press.
NORTH, A. C. T. & PHILLIPS, D. C. (1969). *Prog. Biophys.* **19**, part 1, 1.
RANGO, C. DE, MAUGUEN, Y. & TSOUCARIS, G. (1975). *Acta Cryst.* A**31**, 227–233.
RANGO, C. DE, TSOUCARIS, G., & ZELWER, C. (1974). *Acta Cryst.* A**30**, 342–353.
REEKE, G. N. & LIPSCOMB, W. N. (1969). *Acta Cryst.* B**25**, 2614–2623.
SAYRE, D. (1953). *Acta Cryst.* **5**, 60–65.
SAYRE, D. (1974). *Acta Cryst.* A**30**, 180–184.
TSOUCARIS, G. (1970*a*). *Acta Cryst.* A**26**, 492–499.
TSOUCARIS, G. (1970*b*). *Acta Cryst.* A**26**, 499–501.
WEINZIERL, J. E., EISENBERG, D. & DICKERSON, R. E. (1969). *Acta Cryst.* B**25**, 380–387.
YONATH, A., SMILANSKY, A., MOULT, J. & TRAUB, W. (1973). *Abs. 9th. Int. Cong. Biochem.* Stockholm, p. 120.

Acta Cryst. (1976). A**32**, 292

Multiple Diffraction in Diamond*

BY BEN POST

Polytechnic Institute of New York, Brooklyn, N.Y. 11201, *U.S.A.*

(*Received* 2 *May* 1975; *accepted* 23 *September* 1975)

The 002 and 222 multiple diffraction patterns of diamond, originally recorded by Renninger [*Z. Phys.* (1937). **106**, 141–176.], have been reexamined using high-resolution techniques. Several previously unreported features of these patterns have been observed and are discussed.

Introduction

The first systematic investigation of multiple X-ray diffraction effects in single crystals was carried out by Renninger (1937). He recorded and analyzed the 002 and 222 multiple diffraction patterns of diamond crystals – using Cu $K\alpha$ and Mo $K\alpha$ radiations – in what is now generally regarded as a classic study of the phenomenon.

We have recently calculated the azimuthal angles at which multiple diffraction effects may be observed in 002 and 222 'Renninger patterns' of diamond, recorded with Cu $K\alpha$ radiation. These indicate that several features of crystallographic interest, in addition to those described by Renninger, would be revealed if the patterns were recorded with high-resolution techniques. Results of such an investigation are discussed below.

The geometry and intensities of multiple X-ray diffraction effects in single crystals have been discussed by many investigators in recent years, including: Cole, Chambers & Dunn (1962); Moon & Shull (1964); Zachariasen (1965); Caticha-Ellis (1969) and Prager (1971). An extensive bibliography of the subject is included in a review paper by Terminasov & Tuzov (1964), and more recent references are listed by Post (1975).

Experimental

The experimental arrangement used in this investigation is similar to Renninger's, modified to improve resolution (Fig. 1). The X-ray source was a Cu target tube with an effective focal spot size of 400 × 500 μm at a take-off angle of 4°. A 0·5 mm pinhole at the exit end of a 120 cm evacuated tube between the source and the specimen limited the divergence of the incident beam to 2′ of arc.

Two diamond specimens were used. One was a 1 cm square platelet, 2 mm thick, with [001] normal to the large face. It was optically clear and colorless, and exhibited considerable birefringence when examined between crossed polarizers. The other was roughly octahedral in shape, with triangular (111) faces ap-

* Work supported in part by: Contract F44620-74-C-0065, U.S. Army, Joint Services to the Electronics Program.