Acta Cryst. (1985). A41, 3-17

The Extension and Refinement of the 1·9 Å Spacing Isomorphous Phases to 1·5 Å Spacing in 2Zn Insulin by Determinantal Methods

BY C. DE RANGO, Y. MAUGUEN AND G. TSOUCARIS

Centre Pharmaceutique, Université de Paris, Chatenay-Malabry, Paris

AND E. J. DODSON, G. G. DODSON AND D. J. TAYLOR*

Department of Chemistry, University of York, Heslington, York YO1 5DD, England

(*Received* 15 *September* 1983; *accepted* 9 *May* 1984)

Abstract

A new mathematical description of phase relationships which connects different approaches, both in reciprocal and direct space, is formulated. It leads to the development of a novel algorithm for phase extension and refinement based on a probability function for atomic presence. This function, calculated from the elements of the Karle-Hauptman inverse matrix, is used in an iterative procedure. Various tests have been performed on an idealized set of calculated structure factors for an insulin model structure. The method has been applied to experimental data, F_{obs}, and the isomorphous phases for 2Zn insulin. An assessment of the quality of the phase refinement and calculation has been made by comparison with the crystallographically refined phases.

Introduction

Direct methods, at the present stage, are not powerful enough to allow the *ab initio* determination of protein structures. But they can be used in order to improve and/or extend a set of approximate phases at medium resolution. This extension may lead to a set of phases at higher resolution corresponding to a Fourier series which can provide a better set of atomic coordinates for subsequent refinement. Different mathematical approaches have been developed and used in practical work: the tangent formula method (Hendrickson, Love & Karle, 1973); the Sayre (1972, 1974) equations; the determinantal methods (de Rango, Mauguen & Tsoucaris, 1975; Castellano, Podjarny & Navaza, 1973; Podjarny, Yonath & Traub, 1976; Podjarny & Yonath, 1977; Mauguen, 1979; Podjarny, Schevitz & Sigler, 1981); and the filter type methods (Gassmann, 1976; Raghavan & Tulinsky, 1979; Schevitz, Podjarny, Zwick, Hughes & Sigler, 1981; Collins, 1982).

A brief description of the application of the determinantal methods to actual protein structures has been the object of several communications (Xth Int. Congr. Crystallogr., Amsterdam, 1975; IVth Eur. Crystallogr. Meet., Oxford, 1977; XIth Int. Congr. Crystallogr., Warsaw, 1978; de Rango, Mauguen, Tsoucaris, Dodson, Dodson & Taylor, 1979).

Here, we give a new mathematical description, including the connection between different mathematical approaches: the probability of atomic presence $\tau(r)$ function, the regression method and the forbidden domain theory of von Eller. Then, we describe the algorithms and the practical procedure. Next, we give the results of preliminary tests on an idealized set of data constituted of calculated structure factors for an insulin model structure. Finally, we apply the procedure to actual problems using the insulin crystal data and report a detailed analysis of the results of phase extension of the 1·9 Å isomorphous data to 1·5 Å resolution.

I. Theoretical aspects of the regression equation

(*a*) *The probability of atomic presence functions*

Let us consider a periodic function, $\rho(\mathbf{r})$, whose Fourier coefficients are known up to a certain resolution, R. The Fourier series

$$\rho_0(\mathbf{r}) = \sum_{\mathbf{H}} E_{\mathbf{H}} \exp(-2\pi i \mathbf{H} \cdot \mathbf{r}), \qquad |\mathbf{H}| < 1/R,$$

is an approximation to $\rho(\mathbf{r})$. If the set of Fourier coefficients is the only information we possess on $\rho(\mathbf{r})$, it is well known that $\rho_0(\mathbf{r})$ is the best trigonometric approximation to $\rho(\mathbf{r})$. But if we know something *a priori* about $\rho(\mathbf{r})$ (positivity, atomicity *etc.*), then the problem can be restated in terms of probability theory: we are seeking the best approximation $\rho'(\mathbf{r})$ to $\rho(\mathbf{r})$, given a limited set of Fourier coefficients and the *a priori* information, which corresponds to a minimum for the least-squares expression:

$$\int \{\rho(\mathbf{r}) - \rho'(\mathbf{r})\}^2 \, \mathrm{d}^3 r.$$

When $\rho(\mathbf{r})$ is known to be an atomic structure, the problem can be formulated differently. We are now

* Present address: SERC Daresbury Laboratory, Warrington WA4 4AD, England.

0108-7673/85/010003-15$01.50

interested only in the atomic positions, not merely in the value of the $\rho(\mathbf{r})$ function outside the N atomic positions. More precisely, we are looking for a continuous probability density of atomic presence at the point $\mathbf{r}$, given a limited set of Fourier coefficients and the *a priori* information:

$$p(\mathbf{r}/E\text{'s, information}). \tag{1}$$

Actual mathematical expressions for (1) will be discussed shortly. For the moment we assume that such an expression has been computed in a practical case where low resolution and phase errors do not allow the assignment of individual atoms, or even groups of atoms. Then the question arises how to improve the phases and/or extend them to higher resolution (assuming that the moduli of the structure factors at higher resolution are available). The above probability function provides an immediate answer: the Fourier coefficients of (1) are different from those of $\rho_0(\mathbf{r})$ and may provide better and/or extended phases. These phases, associated with the observed moduli $|E_{\mathbf{H}}|$, will be introduced in a new calculation of (1), and so on, as shown in the following scheme:

$$p(\mathbf{r}|\text{moduli}, \Phi_{\text{initial}}, \text{information}) \xrightarrow{\text{FT}} \Phi_{\text{better}} \tag{2}$$

(FT = Fourier transform). This scheme differs from the classical process involving successive calculations of Fourier series and structure factors, where the $\rho_0(\mathbf{r})$ function is interpreted and modified accordingly by the crystallographer to obtain a function $\rho_{\text{better}}(\mathbf{r})$, whose Fourier coefficients are better and/or extended. Here, such an interpretation is completely avoided. The information allowing the interpretation and improvement of $\rho_0(\mathbf{r})$ in the classical process is now (at least partially) incorporated in the probability of atomic presence, expression (1); the improved phases in each cycle are automatically capitalized in the same expression for the next cycle.

We come now to the mathematical expression of (1). The first historical attempt is perhaps the 'forbidden domain' theory of von Eller (1962, 1964), although the mathematical apparatus does not involve probability theory. The main idea can be summarized as follows: let us 'subtract' one atom, at position $\mathbf{r}$, from a given N-atom structure. If $\mathbf{r}$ coincides with an actual atomic vector $\mathbf{r}_j$ ($j = 1, \ldots, N$), this operation results in an imaginary ($N-1$)-atom structure whose structure factors

$$U'_{\mathbf{H}} = U_{\mathbf{H}} - \frac{1}{N}\exp(2\pi i \mathbf{H}\,.\,\mathbf{r})$$

involve all positive scattering factors. It follows that the Karle-Hauptman (1950) determinant associated with this ($N-1$)-atom structure and denoted ${}^sD_m(\mathbf{r})$ (s for the subtracted atom) is non-negative:

$${}^sD_m(\mathbf{r}) \geq 0. \tag{3}$$

But if $\mathbf{r}$ does not coincide with an actual atomic vector $\mathbf{r}_j$, the subtraction now leads to an electron density function ${}^s\rho(\mathbf{r})$ constituted by $(N+1)$ atoms: the actual N atoms plus a 'negative' atom at $\mathbf{r}$. Then, the corresponding ${}^sD_m$ determinant may become negative.

Next, let us assume that the phases of the structure factors involved in a low-order determinant are known. Even if this limited number of phases is not sufficient to produce an interpretable electron density map, it is still possible that some information can be obtained from von Eller's theory about the actual distribution of atoms in the unit cell, as follows:

$${}^sD_m(\mathbf{r}) = \text{Det}\left\{U_{pq} - \frac{1}{N}\exp[2\pi i(\mathbf{H}_p - \mathbf{H}_q)\,.\,\mathbf{r}]\right\}. \tag{4}$$

The ${}^sD_m$ is computed for each position $\mathbf{r}$ of the subtracted atoms. The value ${}^sD_m(\mathbf{r})$ is plotted: the domains in the unit cell corresponding to negative values of ${}^sD_m(\mathbf{r})$ are the 'forbidden domains' for atomic presence (Fig. 2 in von Eller, 1962).

In other words, the above theory may assign the probability zero or non-zero for the atomic presence at position $\mathbf{r}$. In practice, the allowed domains are not small enough for actual applications, except for very simple structures. However, probability theory can bring further restrictions by introducing the notion of 'most probable domains' for atomic presence. This can be done by using in direct space the probabilistic meaning of the maximum determinant rule instead of the strict algebraic meaning of (3) (de Rango, 1969; Tsoucaris, 1970).

$$D_{m(\text{correct phases})} = \text{maximum}. \tag{5}$$

$${}^sD_{m(\text{correct phases})} = \text{maximum}. \tag{6}$$

We recall that the *a priori* information introduced consists of positivity, atomicity, total number of atoms N, knowledge of unitary (U's) or normalized (E's) structure factors and symmetry.

A general and precise mathematical expression for (1) is to be developed. Such an expression should yield a probability density $p = 0$ in all cases where ${}^sD_m(\mathbf{r}) < 0$. This remark suggests a combination of (3) and (6) which can be used in practical work:

$$p(\mathbf{r}/E\text{'s, information}) \propto \begin{cases} {}^sD_m(\mathbf{r}) & \text{if } {}^sD_m(\mathbf{r}) > 0 \\ 0 & \text{if } {}^sD_m(\mathbf{r}) \leq 0 \end{cases}. \tag{7}$$

We note that the probability expression (7) depends on $\mathbf{r}$ only through the value of the determinant.

We cannot, for the moment, provide a strict justification of (6) and (7). However, there exist several independent indications. Firstly, Lajzerowicz & Lajzerowicz (1966) have shown that ${}^SD_N(\mathbf{r})$ has indeed

the meaning of a presence probability function under specific conditions. Next, qualitative arguments as well as experimental evidence (Tsoucaris, 1981; Taylor, Woolfson & Main, 1978) support the correctness of (5).

The same remarks can be applied to ${}^sD_m$ in (6) as well. Finally, it has been shown algebraically from the maximum determinant rule that ${}^sD_m$ is maximum around the atomic position (Knossow, de Rango, Mauguen, Sarrazin & Tsoucaris, 1977). Other algebraic approaches connected with the strict equality can be mentioned (Goedkoop, 1952),

$$D_m = 0 \quad \text{for } m \geq N+1,$$

which also lead to algorithms allowing the atomic positions to be determined from a limited set of structure factors (Collins, 1978; Navaza & Silva, 1979).

We will now derive from the regression equation theory, which has been developed and applied previously in reciprocal space (de Rango, Tsoucaris & Zelwer, 1974), a new approach which will be shown to be closely connected with (7).

(*b*) *The regression equation: connection with direct space*

It has been shown from the maximum determinant rule (Tsoucaris, 1970; de Rango, Mauguen & Tsoucaris, 1975) that the expected value $\bar{E}_{\mathbf{H}}$ of a normalized structure factor is given by the regression equation:

$$\bar{E}_{\mathbf{H}} = \bar{E}_{\mathbf{L}-\mathbf{H}_q} = -\frac{1}{D_{qq}} \sum_{\substack{p=1 \\ p\neq q}}^{m} D_{pq} E_{\mathbf{L}-\mathbf{H}_p}, \tag{8}$$

where D_{pq} are the elements (row p, column q) of the inverse of a Karle–Hauptman (1950) (K–H) matrix $\|U\|_m$ of order m defined by the m vectors $\mathbf{H}_p$ ($p = 1, \ldots, m$), whose elements are $U_{pq} = U_{\mathbf{H}_p - \mathbf{H}_q}$; the m structure factors $E_{\mathbf{L}-\mathbf{H}_p}$ ($p = 1, \ldots, m$) are the elements of the last row of the K–H matrix $\|U\|_{m+1}$ obtained by adding $\mathbf{L}$ to the set $\{\mathbf{H}_p\}$.

On the other hand, it has also been shown (Knossow *et al.*, 1977) that the 'forbidden domain' function of von Eller, ${}^sD_m(\mathbf{r})$, can be written

$${}^sD_m(\mathbf{r}) = D_m - \frac{1}{N} \sum_{p,q=1}^{m} D_{pq} \exp\left[-2\pi i(\mathbf{H}_p - \mathbf{H}_q) \,.\, \mathbf{r}\right] \tag{9}$$

with $D_m = \det\{\|U\|_m\}$.

Collecting together all the terms such that $\mathbf{k} = \mathbf{H}_p - \mathbf{H}_q$, we can rewrite (9) as

$${}^sD_m(\mathbf{r}) = D_m + \frac{1}{N} \sum_{\mathbf{k}} C_{\mathbf{k}} \exp\left[-2\pi i \mathbf{k} \,.\, \mathbf{r}\right] \tag{10}$$

with $C_{\mathbf{k}}$ defined as

$$C_{\mathbf{k}} = -\sum_{p,q=1}^{m} \delta\{\mathbf{k} - (\mathbf{H}_p - \mathbf{H}_q)\} D_{pq}, \tag{11}$$

where

$$\delta\{\mathbf{k} - (\mathbf{H}_p - \mathbf{H}_q)\} = \begin{cases} 1 & \text{if } \mathbf{H}_p - \mathbf{H}_q = \mathbf{k} \\ 0 & \text{if } \mathbf{H}_p - \mathbf{H}_q \neq \mathbf{k} \end{cases}.$$

Now, starting from (8), make $\mathbf{L} = \mathbf{H} + \mathbf{H}_q$ to get the qth estimate of $E_{\mathbf{H}}$:

$$(\overline{E_{\mathbf{H}}})_q = -\frac{1}{D_{qq}} \sum_{\substack{p=1 \\ p\neq q}}^{m} D_{pq} E_{\mathbf{H}-(\mathbf{H}_p - \mathbf{H}_q)}.$$

Assuming the errors in these estimates to be statistically independent for different values of q, the best global estimate $\langle \bar{E}_{\mathbf{H}} \rangle$ is the weighted average of all the $(\overline{E_{\mathbf{H}}})_q$, each weighted by its inverse variance $1/\sigma_q^2 = D_{qq}$. Therefore,

$$\begin{aligned} \langle \overline{E_{\mathbf{H}}} \rangle &= \left(\sum_{q=1}^{m} D_{qq} \right)^{-1} \sum_{q=1}^{m} D_{qq} (\overline{E_{\mathbf{H}}})_q \\ &= \left(\sum_{q=1}^{m} D_{qq} \right)^{-1} \left\{ -\sum_{\substack{p,q=1 \\ p\neq q}}^{m} D_{pq} E_{\mathbf{H}-(\mathbf{H}_p - \mathbf{H}_q)} \right\}. \end{aligned}$$

Using $C_{\mathbf{k}}$ as defined in (11), we get

$$\langle \overline{E_{\mathbf{H}}} \rangle = \left(\sum_{q=1}^{m} D_{qq} \right)^{-1} \sum_{\mathbf{k}\neq\mathbf{0}} C_{\mathbf{k}} E_{\mathbf{H}-\mathbf{k}}. \tag{12}$$

On the r.h.s. of (12) appears the convolution product of two sets of Fourier coefficients; the corresponding functions in direct space are

$$\begin{aligned} \rho(\mathbf{r}) &= \sum_{\mathbf{H}} E_{\mathbf{H}} \exp\left(-2\pi i \mathbf{H} \,.\, \mathbf{r}\right) \\ \tau(\mathbf{r}) &= \sum_{\mathbf{H}\neq\mathbf{0}} C_{\mathbf{H}} \exp\left(-2\pi i \mathbf{H} \,.\, \mathbf{r}\right). \end{aligned} \tag{13}$$

This last function is, apart from an additive constant and a multiplicative constant, the 'forbidden domain' function of von Eller, ${}^sD_m(\mathbf{r})$. Indeed, from (13) and (10) we get

$$\tau(\mathbf{r}) = N({}^sD_m(\mathbf{r}) - D_m) - C(\mathbf{0}).$$

Therefore, the Fourier transform of (12) leads in direct space to the equivalent equation:

$$\overline{\rho(\mathbf{r})} \propto \tau(\mathbf{r})\rho(\mathbf{r}). \tag{14}$$

This equivalence establishes a link between the purely algebraic approach in direct space of von Eller and the statistical considerations in reciprocal space which are the basis of the regression equation.

Equation (14) suggests that the following iterative scheme can be applied in direct space:

$$\rho^{(i+1)}(\mathbf{r}) \propto \tau(\mathbf{r})\rho^{(i)}(\mathbf{r}), \tag{15}$$

where the observed moduli $|E_{\mathbf{H}}|$ are imposed; or, more precisely,

$$\tau(\mathbf{r})\rho^{(i)}(\mathbf{r}) \xrightarrow{\text{FT}^{-1}} (|E_{\text{calc}}|^{(i+1)}, \Phi_{\text{calc}}^{(i+1)}) \tag{16}$$

$$(|E_{\text{obs}}|, \Phi_{\text{calc}}^{(i+1)}) \xrightarrow{\text{FT}} \rho^{(i+1)}(\mathbf{r}), \tag{17}$$

where $\rho^{(i)}(\mathbf{r})$ is the approximation of the electron density obtained in cycle (i). A similarity appears between the above scheme and that used in the electron density modification methods (Collins, 1975; Gassmann, 1966), which can be formulated as

$$\rho^{(i+1)}(\mathbf{r}) \propto g[\rho^{(i)}(\mathbf{r})]\rho^{(i)}.$$

Note that, starting from such a formula, Gassmann (1976) succeeded in obtaining, inversely, the maximum determinant rule.

II. The computational procedure

The problems arising in the use of the theory described above are amenable to standard techniques and, according to the general scheme developed in § I(*b*), an algorithm has been devised using the product function $\tau(\mathbf{r})\rho(\mathbf{r})$. This algorithm is based upon a fixed function $\tau(\mathbf{r})$ which allows us to exploit the initial information (the phases to low resolution, R_0, and all moduli to higher resolution, R_1). Once this function is determined, cyclic calculations, involving only fast Fourier transformations (FFT), lead to the evaluation of new phases corresponding to higher resolution R_1 according to the following general scheme:

$\{\varphi\}_{R_0}$
Moduli, $\|U\|_m^{-1}$
a priori information
$\Big\} \rightarrow$

FFT

$\rightarrow \tau(\mathbf{r})\rho^{(i)}(\mathbf{r}) \xrightarrow[\text{FFT}^{-1}]{} |E_{\text{cal}}| \exp \varphi_{\text{cal}} \xrightarrow[\text{combination}]{\text{phase}} \{\varphi^{(i+1)}\}_{R_1}.$

The algorithm is developed in two main stages.

First stage. A unique $\|U\|_m$ matrix is built up to resolution R_0 and inverted to provide the C's (11), and finally the $\tau(\mathbf{r})$ function from (13).

Second stage. The phases to resolution R_0 are introduced in the $\tau(\mathbf{r})\rho(\mathbf{r})$ expression; operation (16) yields better and/or extended phases $\varphi^{(i+1)}$, leading by (17) to a better $\rho^{(i+1)}(\mathbf{r})$ function which is recycled in (16), and so on.

The principle of this procedure is similar to that used in the application of the regression equation method in reciprocal space, which was described in a previous paper (de Rango, Mauguen & Tsoucaris, 1975), referred to hereafter as paper 1. However, in spite of the formal mathematical equivalence between the regression equation and the $\tau(\mathbf{r})\rho(\mathbf{r})$ function, the practical algorithms of the second stage are different. We recall that in reciprocal space, once the matrix inversion is done, the cycling procedure consists of using a set of structure factors $E_{\mathbf{L}}$ to build up the last row and column of a set of matrices $\|U\|_{m+1}$ and then calculating for each element of these rows an estimated phase. Finally, the partial determinations obtained by all the $\|U\|_{m+1}$ matrices are combined for each reflexion.

The development of the computational procedures was carried out using calculated structure factors derived from a set of atomic coordinates for 2Zn pig insulin in order to avoid experimental errors while investigating the ability of the method to extend phases with reasonable accuracy and to produce interpretable electron density maps. In all stages, we use normalized structure factors calculated by the Wilson method.

(*a*) *The fixed* $\tau(\mathbf{r})$ *function*

The matrix construction at this stage is difficult to control as there is no general procedure that will always produce a 'good matrix'. The main problem is that the phases of the $m(m-1)/2$ structure factors, elements of an m-order matrix, may be largely erroneous or even totally unknown. These structure factors depend in fact only on the choice of the $(m-1)$ elements of the top row ('basic element'). It has been shown, in paper 1, that a selection must be applied in order to get good conditions in the phase refinement and extension procedure. This selection becomes too restrictive if structure factors which generate incomplete rows are not accepted as basic elements. But we can reasonably expect that a certain tolerance is acceptable in the number of unknown elements. An *a priori* estimation of the occupancy factor (percentage of known elements) can be achieved (Mauguen, 1979), as follows.

Suppose that the structure factors, which are known in modulus and phase, are included in a sphere of radius R, and that the basic elements $U_{\mathbf{H}_j}$ are chosen in a sphere of radius r $(r < R)$. It is clear that all elements $U_{\mathbf{H}_i-\mathbf{H}_j}$ will be known if the elements $U_{\mathbf{H}_i}$ are chosen inside the sphere of radius r with $r < R/2$. For values of r higher than $R/2$, the occupancy factor (OF) will decrease when r increases. The variation of the OF can be estimated *a priori* as a function of the values of $x = r/R$ (see Fig. 1 and Appendix I). The

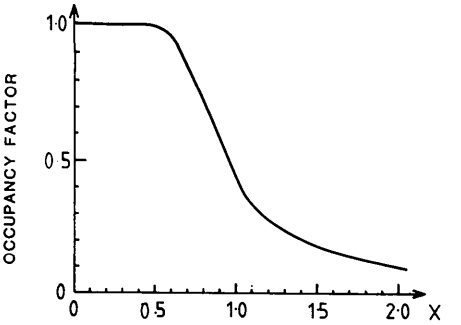

Fig. 1. Variation of the occupancy factor (OF) as a function of $x = r/R$.

curve is very close to the experimental one given by Podjarny & Yonath (1977). From practical applications it was found that good phase estimations can be obtained for values of OF higher than 0·7 corresponding to a ratio $r/R<0{\cdot}8$. But in practice the basic elements are selected by using several criteria which maximize the value of the modulus of the elements and reduce the number of the unknown elements in each row. Under these conditions the values of the OF are higher than 0·7 even for values of x higher than 0·8. This fact is very important since it gives rise to a larger choice of basic elements (if $x=0{\cdot}8$ instead of 0·5, the number of possible basic elements is multiplied by 4).

The number of unknown elements in the matrix can be reduced by evaluating approximately the phases of the unknown elements $U_{\mathbf{H}_i-\mathbf{H}_j}$, either by symbolic addition involving the phases of the two basic elements $U_{\mathbf{H}_i}$ and $U_{\mathbf{H}_j}$

$$\varphi_{\mathbf{H}_i-\mathbf{H}_j}=\varphi_{\mathbf{H}_i}-\varphi_{\mathbf{H}_j},$$

or from the classical tangent formula using the structure factors included in two rows

$$\varphi_{\mathbf{H}_i-\mathbf{H}_j}=\text{phase of}\left(\sum_p E_{\mathbf{H}_i-\mathbf{H}_p}E_{\mathbf{H}_p-\mathbf{H}_j}\right).$$

However, the phases then obtained can be largely erroneous for protein structures and lead to negative determinants. In practice, if the OF is large enough, the most efficient treatment is to put the elements $U_{\mathbf{H}_i-\mathbf{H}_j}$ equal to zero. Moreover, when a certain number of off-diagonal elements of a matrix $\|U\|_m$, considered as a covariance matrix, are replaced by zero we get an approximation in which we assume that the correlation coefficient between two random variables, E_i and E_j, is equal to zero. This means that

$$\langle E_i^* E_j\rangle=\langle E^*_{\mathbf{L}-\mathbf{H}_i}E_{\mathbf{L}-\mathbf{H}_j}\rangle\mathbf{L}=0$$

instead of

$$\langle E_i^* E_j\rangle=U_{\mathbf{H}_i-\mathbf{H}_j}.$$

This has also been discussed by Castellano *et al.* (1973) and applied by Podjarny & Yonath (1977). For practical applications, the unobserved structure factors and those having a phase determined with too low a figure of merit (usually lower than 0·3) have been treated in the same way.

Apart from these procedures, the algorithm for building up the matrix was developed following the guidelines given in paper 1.

Firstly, a certain number of reflexions which will constitute the set of possible basic elements are chosen as a function of the modulus and the figure of merit (usually FOM > 0·7). These reflexions lie in a resolution range such that a sufficient number of terms of the matrix have a resolution higher than the initial limit of the input of the matrix. Symmetrical reflexions may be included, but in practice it has been found desirable to restrict the elements of the top row to a unique set in a fixed order (usually 50% of the final order required). Very-low-resolution reflexions are also omitted as they lead to the introduction of structure invariants into the determinant whose probabilities are estimated too high when calculated on the basis of the magnitude $|E_{\text{obs}}|$.

Table 1. *Parameters used in the construction of matrices* $\|U\|_m$ *in different cases of phase refinement and extension*

(1) From 3·5 to 2·5 Å (calculated data).
(2) From 2·8 to 1·9 Å (experimental data).
(3) From 1·9 to 1·5 Å (experimental data).

	Selection of basic elements				Value of
	Order	Resolution d	$\|E\|_{\min}$	OF	D_m
1	300	$4{\cdot}2\le d\le 9$ Å	1·1	0·93	$0{\cdot}4\times10^{-17}$
2	400	$4{\cdot}0\le d\le 9$ Å	1·2	0·92	$0{\cdot}4\times10^{-22}$
3	400	$2{\cdot}5\le d\le 6$ Å	1·4	0·88	$0{\cdot}3\times10^{-17}$

Next, the first element is chosen. Usually, the highest modulus reflexion is taken but any other reflexion can be imposed. Then, for a given order m_0, all the interactions $U_{\mathbf{H}_i-\mathbf{H}_j}$, between the (m_0-1) basic elements already chosen and each element of the set of possible basic elements are sorted. Among all possible basic elements the next one selected corresponds to the maximum of the following expression:

$$C_1=\sum_p |E_{\mathbf{H}_i}E_{\mathbf{H}_p}E_{\mathbf{H}_i-\mathbf{H}_p}|\cos(\varphi_{\mathbf{H}_p}-\varphi_{\mathbf{H}_i}+\varphi_{\mathbf{H}_i-\mathbf{H}_p}).$$

The summation extends over the basic elements already chosen. The mean value of C_1 for each row should decrease smoothly with increasing order and negative values occur rarely, if at all, at high orders. Additional criteria are introduced: for each row the mean value of the figure of merit and the occupancy factor are kept higher than given limits. This step is repeated until the required matrix order is obtained and until there are no more possible basic elements.

Table 1 gives, as an example, the parameters used in the construction of several matrices with different ranges of resolution data. These matrices have been used effectively in phase determination procedures and they all led to satisfactory results, as will be shown below. However, it should be noted that for low-resolution data (3·5–2·5 Å) the choice of the parameters may be very critical in building up the matrix since it can happen that the phase determination procedure diverges even if the determinant and all eigenvalues are positive.

The inversion procedure makes use of the Cholevski method, which permits us to take into account the advantage of the Hermitian properties of the K–H matrices and leads to a very easy file treatment without requiring the storage of the full matrix in the core memory of the computer. The K–H matrices used are usually well conditioned (the off-diagonal elements are small with respect to the

diagonal elements), providing the matrix order is not too near that for which the matrix becomes singular. For example, this property allows the calculation of the elements D_{pq} of the inverse matrix in single precision.

Finally, the quality of the matrix can be assessed by the following criterion: the value of the determinant should be as low as possible (Appendix II) and should not oscillate as it decreases in value with increasing order. The Fourier coefficients of the $\tau(\mathbf{r})$ function are directly determined from the values of D_{pq} using the relation (11).

(b) The product $\tau(\mathbf{r})\rho(\mathbf{r})$ function

The method developed in paper 1 presented two main disadvantages: it was difficult to control *a priori* the number of estimated phases since they depend mainly on the choice of the structure factors E_L, and it required a very large computing time (proportional to N^2). The direct-space formulation makes it possible to overcome these difficulties: it leads to a systematic determination of all structure factors and to a computing time which is considerably reduced, since the application of formula (14) makes implicit use of mM of regression equations (8) (M total number of reflexions), which is much larger than the number that can be exploited effectively in reciprocal space. The calculation can be done easily by using only three Fourier transformations; the computing time is then proportional to $N \log N$ instead of N^2.* This scheme, however, may present a drawback. In reciprocal space, the structure factors E_L were selected as a function of several criteria such as occupancy factor and mean value of the moduli of the elements for each vector E_L. Here, no such selection is possible. Nevertheless, this is largely compensated for by the greater number of implicit estimations.

Preliminary tests of the procedure were carried out by calculating phase extensions at several ranges of resolution and examining the phase errors as a function of resolution and $|E|$ (moduli of normalized structure factors). The results indicate that the product $\tau(\mathbf{r})\rho(\mathbf{r})$ remains reasonably stable and the algorithm does extend phase information with reasonable precision. The results obtained in a phase extension from 2·3 to 1·9 Å are summarized in Table 2. The phases are stable after two cycles of calculations, indicating that only a small number of cycles will be necessary in practical applications. Table 3 gives the mean values of the phase errors under two different conditions. In each case the mean value of phase errors is low whatever the range of the moduli.

Table 2. *Mean values of phase errors for the first two cycles of phase extension calculations from* 2·3 *to* 1·9 Å *resolution*

Ref: refinement range; Ext: extension range.

	Number of determined phases	$\|E\|$	Resolution (Å)	Mean value of phase errors Cycle 1	Cycle 2
Ref.	3833	>0·0	$d > 2·3$	33·1	32·2
Ext.	2977		$1·9 < d < 2·3$	40·7	40·7
Ref.	1489	>1·0	$d > 2·3$	16·7	13·7
Ext.	1160		$1·9 < d < 2·3$	23·4	23·7
Ref.	136	>2·0	$d > 2·3$	10·7	8·0
Ext.	77		$1·9 < d < 2·3$	13·3	12·8

Table 3. *Mean values of phase errors in*: (*a*) *phase refinement and extension from* 2·8 *to* 1·9 Å; (*b*) *phase refinement and extension from* 3·5 *to* 2·5 Å

Ref: refinement range; Ext: extension range.

	Number of determined phases	$\|E\|$	Resolution (Å)	Mean value of phase errors
(*a*)				
Ref.	2165	>0·1	$d > 2·8$	34
Ext.	4346		$1·9 < d < 2·8$	52
Ref.	912	>1	$d > 2·8$	19
Ext.	1738		$1·9 < d < 2·8$	36
Ref.	111	>2	$d > 2·8$	12
Ext.	102		$1·9 < d < 2·8$	22
(*b*)				
Ref.	1091	>0·1	$d > 3.5$	35
Ext.	1904		$2·5 < d < 3·5$	56
Ref.	565	>1	$d > 3·5$	21
Ext.	628		$2·5 < d < 3·5$	37
Ref.	96	>2	$d > 3·5$	15
Ext.	24		$2·5 < d < 3·5$	11

However, at high or medium resolution, the phase errors of the extension region are very similar to those of the refinement region; but when the resolution of the data is lower the phase error is less for the refined phases than for the extended phases.

As a further test, the validity of the procedure was checked to higher resolution, with an order-400 matrix constructed using both moduli and phases, to 1·5 Å resolution, calculated from the refined atomic coordinates and associated temperature factors. The results showed that, after two cycles of refinement, the average error for $|E| > 1$ was only about 25° and that no pronounced divergence from the refined structure had occurred.

Figs. 2(*a*) and (*b*) illustrate these results by giving, for different limits of modulus, the mean values of phase errors as a function of the resolution for the first and second cycles, respectively; they are nearly constant. It is also clear that they are lower for the highest moduli.

It was found that, for practical applications, the best method of recycling was to include in the next

* CPU time on IBM 370/168: 2 h 30 min per cycle for 2Zn insulin. The CPU time for the matrix inversion is 10 min; despite being proportional to m^3 it is not the limiting step in the procedure. Moreover, in the direct-space procedure used here, $\tau(r)$ is calculated only once.

estimate of $\rho^{i+1}(\mathbf{r})$ all the $|E| > 1{\cdot}0$ while progressively increasing resolution until the resolution limit of the moduli was reached. This is justified by the fact that the precision is best for the structure factors of highest modulus.

(*c*) *Phase combination*

In each following cycle, or in the construction of the electron density maps, it is desirable to associate a figure of merit with each structure factor according to the accuracy of the calculated phases. This has been performed using a weighting scheme based on the modulus of the $E_{\mathbf{h},\mathrm{calc}}$ coefficients from the Fourier transform of the product $\tau(\mathbf{r})\rho(\mathbf{r})$, following a procedure which was developed by Bricogne (1976). For more efficiency, the information included in the results may be combined with that which comes initially from multiple isomorphous replacement or anomalous scattering methods. This problem has been considered previously by several authors (Woolfson, 1956; Sim, 1960).

Blow & Crick (1959) have shown that the most probable phase is not necessarily the best to use, and that to obtain the minimal r.m.s. error of the electron density the optimal phase value $\hat{\varphi}_h$ and the associated weight m (figure of merit) of a structure factor $E_{\mathbf{h},\mathrm{obs}}$ must be defined by

$$m|E_{\mathbf{h},\mathrm{obs}}| \exp(i\hat{\varphi}_{\mathbf{h}}) = \int_0^{2\pi} p(\varphi_{\mathbf{h}})|E_{\mathbf{h},\mathrm{obs}}| \exp(i\varphi_{\mathbf{h}})\, \mathrm{d}\varphi_{\mathbf{h}},$$

where $p(\varphi_h)$ is the probability density of the phase φ_h. The figure of merit is the mean value of the phase error cosine.

The phase information may come from several sources, but in this treatment we limit it to the case where it is provided only by isomorphous replacement (initial information to medium resolution) and by direct methods (final information to high resolution).

The probability density of the phases associated with different information has been studied in the case of protein structures (Rossman & Blow, 1961; Hendrickson & Lattman, 1970). The main point to emphasize is that the function can be written as follows:

$$p_{\mathrm{iso}}(\varphi_h) \propto \exp\{A_{\mathbf{h},\mathrm{iso}} \cos\varphi_{\mathbf{h}} + B_{\mathbf{h},\mathrm{iso}} \sin\varphi_{\mathbf{h}} + C_{\mathbf{h},\mathrm{iso}} \cos 2\varphi_{\mathbf{h}} + D_{\mathbf{h},\mathrm{iso}} \sin 2\varphi_{\mathbf{h}}\}.$$

In practical applications this approach makes it simple to handle the contributions to the phase from different sources, *e.g.* partial structures, non-crystallographic symmetry direct methods (Bricogne, 1976; Hendrickson, Love & Karle, 1973). It leads to an addition of the coefficients, each of which corresponds to the different probability densities.

In the Hendrickson & Lattman procedure, the coefficients $A_{\mathbf{h},\mathrm{iso}}$; $B_{\mathbf{h},\mathrm{iso}}$; $C_{\mathbf{h},\mathrm{iso}}$; $D_{\mathbf{h},\mathrm{iso}}$ came directly from the heavy-atom refinement in the isomorphous derivatives through the lack of closure defined by the intensities. In the application we are concerned with here, the figures of merit associated with the structure factors were also available. From these two quantities we can recalculate the values of $A_{\mathbf{h},\mathrm{iso}}$ and $B_{\mathbf{h},\mathrm{iso}}$, if we suppose that the density probability is unimodal instead of bimodal. This approximation is justified by the very good quality of the isomorphous phases.

When we calculate the Fourier transform of the product function $\tau(\mathbf{r})\rho(\mathbf{r})$, we get an estimation of the phase ($\varphi_{\mathbf{h},\mathrm{calc}}$) and the modulus $|E_{\mathbf{h},\mathrm{calc}}|$ of each structure factor. Following a procedure developed first by Bricogne (1976), the values of the $E_{\mathbf{h},\mathrm{calc}}$'s can be used to derive for each structure factor, $E_{\mathbf{h},\mathrm{obs}}$, a probability function of the associated phase defined by

$$p_{\mathrm{calc}}(\varphi_{\mathbf{h}}) = \exp A_{\mathbf{h}} \cos(\varphi_{\mathbf{h}} - \varphi_{\mathbf{h},\mathrm{calc}})$$

with

$$A_{\mathbf{h}} = 2|E_{\mathbf{h},\mathrm{calc}}|\,|E_{\mathbf{h},\mathrm{obs}}|/k,$$

where k is obtained from the mean discrepancy between $|E_{\mathbf{h},\mathrm{calc}}|^2$ and $|E_{\mathbf{h},\mathrm{obs}}|^2$ in ranges of $\sin^2\theta/\lambda^2$:

$$k = (\sum ||E_{\mathbf{h},\mathrm{obs}}|^2 - |E_{\mathbf{h},\mathrm{calc}}|^2|)/M_k$$

and M_k is the number of reflexions included in the summation.

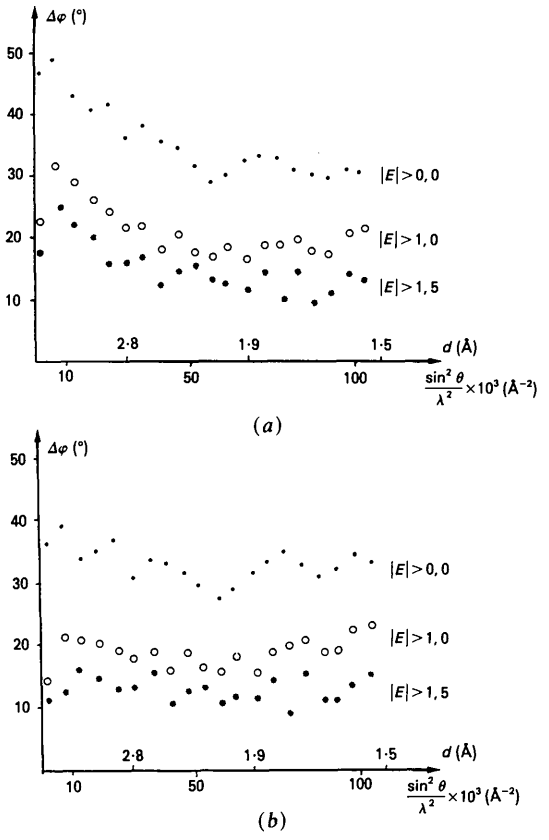

Fig. 2. Distribution of phase errors as a function of resolution for different limits of moduli. Phase refinement to 1·5 Å. (*a*) Cycle 1; (*b*) cycle 2.

The optimal value φ_h of the phase is also the most probable and the associated figure of merit is

$$m = I_1(A_h)/I_0(A_h).$$

The final probability of the phase can be defined by

$$p_f(\varphi_h) = p_{iso}(\varphi_h)p_{calc}(\varphi_h),$$

the product of the probability densities derived respectively from the results of direct and isomorphous replacement methods. This leads us to consider the two sets of information as independent. It is clear that this hypothesis is not exactly justified because the isomorphous phases are used to initiate the direct-method procedures. However, it appears very difficult, even impossible, to account for the dependency rigorously. We must also note that the final probability function does not allow for the experimental errors (inaccuracy of intensities, normalization of the structure factors) and the fact that the phases are initially supposed known.

We made use of this weighting procedure in the electron density calculations but not in the cyclic procedure. However, it would be possible to introduce it there, which would improve convergence.

III. Application to 2Zn insulin diffraction data

The effectiveness of the matrix methods in phase refinement and extension has been successfully demonstrated with model data calculated for 2Zn pig insulin which provide a good check on the phases and electron density at *all resolutions.*

Data for the next part of the analysis consisted of 13 771 structure amplitudes to 1·5 Å resolution as measured in Oxford (Dodson, Dodson, Hodgkin, Isaacs & Vijayan, 1975). The magnitudes of E's appearing in the phase refinement and extension were determined by placing the data set on an absolute scale and normalizing it by taking into account all the atoms of the protein molecule and 285 oxygen atoms of the solvent.

A set of 6700 phases to 1·9 Å resolution produced by the multiple isomorphous replacement method were used as a starting phase set. Phase refinement and extension were performed from 1·9 to 1·5 Å with 400-order matrices generated from observed moduli to 1·5 Å.

To assess the quality of the results, we made use of the final set of atomic coordinates produced by refinement to 1·5 Å resolution with standard techniques (Dodson, Dodson, Hodgkin & Reynolds, 1979), and we compared the phases produced by matrix methods, as well as the corresponding electron density maps, with those calculated from the refined atomic coordinates and associated temperature factors.

(*a*) *Regression formula*

In the preliminary application of the regression equation method in reciprocal space, given in paper I, the influence of the matrix order and the data resolution on the accuracy of the results have been investigated. Two trials have been performed, *viz* phase extensions from 2·8 to 2·2 Å and from 1·9 to 1·5 Å.

For the 2·8 to 2·2 Å phase extension, the quality of the electron density maps constructed by using the refined and extended phases showed a clear improvement with respect to those derived from the set of 2·8 Å isomorphous phases. However, a precise analysis was not carried out owing to the progress with the 1·5 Å study.

In the case of the 1·9 to 1·5 Å phase extension, the results are reviewed in detail. The parameters used in the construction of the matrices are given in Table 1. A new matrix was built up after each cycle since the increasing number of determined phases led to a larger number of possible basic elements which resulted in a higher mean value for the modulus of the matrix elements. The results obtained in the phase determination procedure are summarized in Table 4. The angle shifts, which decrease quickly in three cycles, are given as a function of the figure of merit in Table 5. A comparison between the phases calculated by phase extension and the phases (φ_{ref}) of the refined structure illustrate the effectiveness of this method (Table 6):

the phases out to 1·9 Å which were initially determined by the isomorphous replacement method are clearly improved, particularly for structure factors with large moduli;

a large number of new phases (1·9–1·5 Å) are estimated with a good accuracy;

the angle shifts are very low after two cycles of calculations.

The improvement of electron density maps with respect to the 1·9 Å isomorphous phase maps is evident (Figs. 3–7). A rapid and precise plotting of the coordinates of 755 atoms was achieved very easily. They were used in seven cycles of difference Fourier synthesis calculations, followed by regularization (Dodson, Isaacs & Rollett, 1976) during which the R factor decreased from 0·40 to 0·27. More than 107 independent water molecules were identified. Fig. 3 gives an example of the corresponding electron density maps.

(*b*) *Direct-space procedure*

A detailed analysis of the results obtained in phase refinement and extension from 1·9 to 1·5 Å resolution is given as an example of the direct-space formulation procedure using the $\tau(\mathbf{r})\rho(\mathbf{r})$ expression. The Fourier coefficients C_H of the $\tau(\mathbf{r})$ function were derived from an order-400 matrix which was constructed with parameters given in Table 1. The refined and extended

Table 4. *Selection of* $\|U\|_{m+1}$ *matrices*

L: Miller indices of the basic element for the last column.

	Number of determined phases	⟨FOM⟩	Mean values of phase shifts (°) Before phase combination	After phase combination	Number of matrices $\|U\|_{m+1}$	Selection of L Resolution d	$\|E_L\|_{min}$
Isomorphous phases	6325	67					
1st cycle	7233	74	74	39	1000	$3{\cdot}5 \le d \le 5{\cdot}5$ Å	1·55
2nd cycle	9148	76	43	19	750	$2{\cdot}3 \le d \le 4$ Å	1·35
3rd cycle	10417	80	16	8	750	$2{\cdot}1 \le d \le 3{\cdot}3$ Å	1·40

Table 5. *Mean values of angle shifts as a function of figure of merit*

FOM	Cycle 1 $\Delta\Phi_{reg}$ (°)	Cycle 1 $\Delta\Phi$ after phase combination (°)	Cycle 1 Number of determined phases	Cycle 2 $\Delta\Phi_{reg}$ (°)	Cycle 2 $\Delta\Phi$ after phase combination (°)	Cycle 2 Number of determined phases	Cycle 3 $\Delta\Phi_{reg}$ (°)	Cycle 3 $\Delta\Phi$ after phase combination (°)	Cycle 3 Number of determined phases
0·0-0·5	85	75	819	64	53	364	22	17	572
0·5-0·9	71	40	2065	48	23	2626	19	9	3082
0·9-1·0	56	11	763	31	9	1317	12	4	2493

Table 6. *Mean values of the differences between the phases calculated from the refined atomic coordinates and those calculated by the regression equation method* (°)

Resolution (Å): $d>$	$\|E\|>0{\cdot}1$ 1·5	$\|E\|>0{\cdot}1$ 1·9	$\|E\|>1$ 1·5	$\|E\|>1$ 1·9	$\|E\|>1{\cdot}5$ 1·5	$\|E\|>1{\cdot}5$ 1·9
Isomorphous phases		60·0 (6372)		52·8 (2345)		53·9 (616)
1st cycle	58·6 (7253)	56·0 (6444)	49·5 (2939)	43·0 (2389)	43·1 (723)	39·2 (623)
2nd cycle	59·0 (9153)	54·0 (6444)	51·6 (4180)	38·9 (2389)	42·4 (994)	34·6 (623)
3rd cycle	62·6 (10412)	54·1 (6444)	57·3 (4973)	39·2 (2389)	48·8 (1158)	34·7 (623)

phases were determined directly by five cycles of calculations with the same function $\tau(\mathbf{r})$.

Phase results

The quality of the results is illustrated in Fig. 4 which gives for structure factors of modulus higher than 1 the mean values of the differences between the phases φ_{ref}, derived from the atomic coordinates refinement, and the phases φ_{ext}, calculated by the extension phase method, as a function of resolution,

$$\Delta\varphi_1 = \langle|\varphi_{ref} - \varphi_{ext}|\rangle_{\mathbf{H}}.$$

These values are to be compared with the differences $\Delta\varphi_2$ obtained, in the same conditions, for the phases φ_{iso} determined by the isomorphous replacement method to 1·9 Å resolution,

$$\Delta\varphi_2 = \langle|\varphi_{ref} - \varphi_{iso}|\rangle_{\mathbf{H}}.$$

We recall that we made use of the phases φ_{iso} as starting set. Fig. 4 shows that not only are the phases φ_{iso} corresponding to 1·9 Å resolution improved but a large number of new phases (1·9–1·5 Å) are determined with a good accuracy. For 5050 structure factors, with a modulus higher than 1 and to 1·5 Å resolution , $\Delta\varphi_1$ is 38°, where $\Delta\varphi_2$ is 54° for 2090 structure factors ($|E|>1$ and 1·9 Å resolution).

Table 7 gives the values of $\Delta\varphi_1$ for different ranges in modulus and resolution. All the structure-factor phases (13 337) are determined with a phase error of 52°. This means that, for the doubled number of

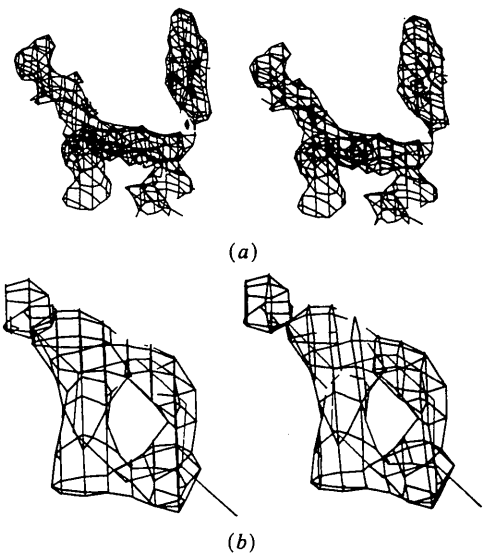

Fig. 3. Electron density after phase refinement and extension from 1·9 to 1·5 Å by regression equation method. (*a*) Molecule 1, residues A13–A14; (*b*) molecule 1, tyrosine A14.

Table 7. *Phase extension from* 1·9 *to* 1·5 Å *mean phase errors as a function of resolution and moduli of normalized structure factors*

Ref: refinement range; Ext: extension range.

	Number of determined phases	$\|E\|$	Resolution (Å) $d>$	$\Delta\varphi$ (°) $\Delta\varphi_2$	$\Delta\varphi_1$
Ref.	600	>1·5	1·9	54	27
Ref. + Ext.	810	>1·5	1·7		27
Ref. + Ext.	1100	>1·5	1·5		28
Ref.	2390	>1·0	1·9	53	33
Ref. + Ext.	3550	>1·0	1·7		35
Ref. + Ext.	5050	>1·0	1·5		38
Ref.	6500	>0·1	1·9	61	49
Ref. + Ext.	10 200	>0·1	1·7		51
Ref. + Ext.	13 330	>0·1	1·5		52

structure factors to 1·5 Å resolution, the precision is better than that of the phases φ_{iso} at 1·9 Å resolution. As was also the case for the preliminary tests, the phase errors are lower for the highest moduli and do not depend very much on the resolution.

The variation of the phase errors is given in Table 8. These values stabilize very quickly and only three cycles of calculations are sufficient. The angle shifts for the two last cycles is 10°. The figure of merit m_{exp}, calculated following the scheme mentioned in § II(*c*), gives a satisfactory estimation of the precision, as is illustrated by the curve in Fig. 5, which represents the variations of $\Delta\varphi_1$ as a function of m_{exp}. For the electron density calculations the different information for the phases φ_{ext} and φ_{iso} has been combined following the procedure described above.

Another attempt to refine and extend phases has been performed in the case of medium resolution (2·8 to 1·9 Å). The results have shown that it was possible to get a reasonably good phase determination after three cycles of calculations. The mean value $\Delta\varphi_3$ of the difference between the extended phases φ_{ext} and the refined phases φ_{ref} is equal to 47°. The error in the calculated phases is smaller for the refinement region (up to 2·8 Å) than for the extension region (2·8 to 1·9 Å) but, in this last case, the error is nearly constant as a function of resolution.

Some features of the agreement between the different phase-refinement-calculated phases and the crystallographically refined phases deserve comment. Firstly, the improvement in the phases (where $E>1·0$) is distinct and for the 1·5 Å data the improvements extend to the data limit. Secondly, the 1·9 Å extension works less well; the improvement in this study is only convincing out to 2·6 Å spacing. Thirdly, the crystallographically refined phases agree better with the phases extended to 1·9 Å from 2·8 Å but at low resolution agree better with the isomorphous phases. This suggests that the phase refinement of the low-angle terms, which contain important and often large contributions from solvent and disordered atoms, suffers by inclusion in the 1·9–1·5 Å data, which of course contain scattering only from well defined atoms.

Electron density maps

A detailed analysis of the electron density maps calculated at 1·5 Å resolution after refinement and extension of phases has been performed. A weighting scheme has been applied following the procedure given in § II(*c*).

The overall impression is that the maps correspond to a resolution of almost 1·5 Å (Figs. 6, 7*b*, 8*b*). They show a distinct improvement with respect to the *initial* electron density maps, *i.e.* calculated with the isomorphous phases at 1·9 Å resolution, and are, in a large part, very close to the *final* maps, *i.e.* calculated after the refinement of the atomic coordinates at 1·5 Å

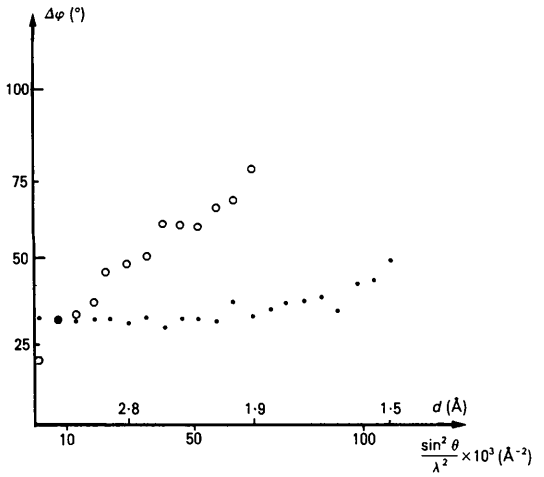

Fig. 4. Distribution of $\Delta\varphi_1$(●) and $\Delta\varphi_2$(○) as a function of resolution. $\Delta\varphi_1$: errors in the phases after phase refinement and extension from 1·9 to 1·5 Å resolution. $\Delta\varphi_2$: errors in the initial phases (1·9 Å).

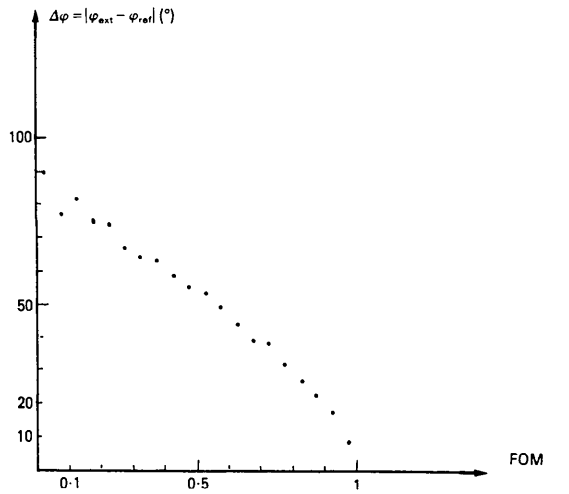

Fig. 5. Distribution of phase errors calculated as a function of figure of merit before combination with isomorphous phases.

Table 8. *Variation of the phase errors during the five cycles of calculations*

	$\|E\|>0\cdot1$			$\|E\|>1\cdot0$			$\|E\|>1\cdot5$		
Resolution (Å) $d>$	1·5	1·7	1·9	1·5	1·7	1·9	1·5	1·7	1·9
Cycle 1 $\Delta\varphi_1$ (°)	61·4	58·3	53·7	48·9	43·9	38·6	39·6	34·6	33·1
Cycle 2 $\Delta\varphi_1$ (°)	55·3	52·9	48·9	42·3	37·9	34·6	33·3	31·0	31·0
Cycle 3 $\Delta\varphi_1$ (°)	52·9	51·2	48·9	39·3	35·4	32·8	28·9	27·2	26·7
Cycle 4 $\Delta\varphi_1$ (°)	51·9	50·6	48·4	37·6	34·1	32·2	27·8	26·0	25·9
Cycle 5 $\Delta\varphi_1$ (°)	52·3	51·0	49·0	38·3	35·0	33·0	28·6	27·3	26·8

(Figs. 7*c*, 8*c*). However, the comparison is less favourable for the regions of low-level electron density corresponding to the disordered atoms of the side chains and to solvent molecules having a high temperature factor, which are also less accurately positioned in the final maps.

We discuss below several particular features.

Disulphide bridges. The six disulphide bonds of the asymmetric unit cell were already almost resolved in the initial maps. The position of all of them were well defined but, compared to the final position, some were in error by more than 0·5 Å. Now, they all come out very clearly in the highest levels of the electron density. Fig. 6 gives an illustration of one of them: cystine A20-B19 mol. 2. The *S* atoms move towards their correct positions which can be determined with good accuracy. The corresponding atomic coordinates are better than those derived from the initial maps. The deviation is now about 0·2 Å except for one of them where the discrepancy is still 0·4 Å.

Polypeptide chain. As is the case for the electron density in the 1·9 Å map, the polypeptide chain is well defined. The atoms are now very close to being individually resolved and, in these regions, the coordinates can be placed with accuracy, as is illustrated in Figs. 6 and 7(*b*).

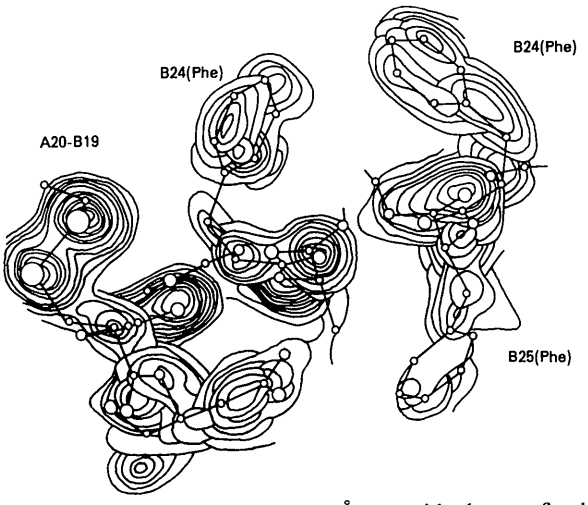

Fig. 6. The electron density in the 1·5 Å map with phases refined and extended from 1·9 Å. There is a satisfactory correspondence between the electron density maxima and the refined atomic position. The height of the density also corresponds well to the atomic number and the definition of the atoms. Note that the buried residues A20–B19 cystine, B24 phenylalanine have well defined strong density; B25 phenylalanine on the surface is a mobile sidechain with the weak density expected. The contour levels are approximately 0·3 e $Å^{-3}$ beginning at 0·3 e $Å^{-3}$.

Side chains. Side chains with small thermal motions were clearly described in the initial maps at 1·9 Å. Now they all have been further improved and they appear easily interpretable in regions for which the density is high. This is the case for the aromatic rings of phenylalanine, B24 of mols. 1 and 2 (Fig. 6) situated at the interfaces between the two monomers which constitute the asymmetric unit cell. For these two residues the temperature factor is about 10 $Å^2$. Side chains having higher thermal vibration appear in regions of lower density and are less accurately defined. However, it is worth noting that even these residues are often still improved after performing phase refinement and extension.

Spurious peaks due to experimental and phase errors were present in the initial maps and sometimes made the interpretation of the corresponding regions difficult. After phase extension, these peaks are often reduced, signalling unreliable density, and making any atomic assignment risky at this stage of the structure's interpretation. Fig. 8 gives an illustration of one of these areas in the different electron density maps. The final interpretation, derived from the refinement of atomic coordinates, is quite different from the initial interpretation at 1·9 Å resolution.

Zn *atoms.* The rhombohedral form of insulin is obtained by crystallization in the presence of zinc, and the asymmetric unit cell contains two Zn atoms, placed in special positions along the threefold axis. The initial maps were very poorly defined around the positions of these atoms. The interpretation of these maps led to placing the two atoms at a half unit cell apart along the threefold axis. This corresponds to an error of about 1·5 Å. For this reason, the interpretation of the structure of the solvent molecules was very difficult in this area. The position of the Zn atoms derived from the maps calculated after phase extension are clearly closer to the final maps although the error is still about 0·7 Å. However, the improvement in the density allows easier interpretation of the solvent features in this area.

Solvent molecules. A general characteristic of many direct methods, especially those using the tangent formula, is to flatten further the areas corresponding to low levels of electron density, making their interpretation sometimes difficult, even impossible. It is

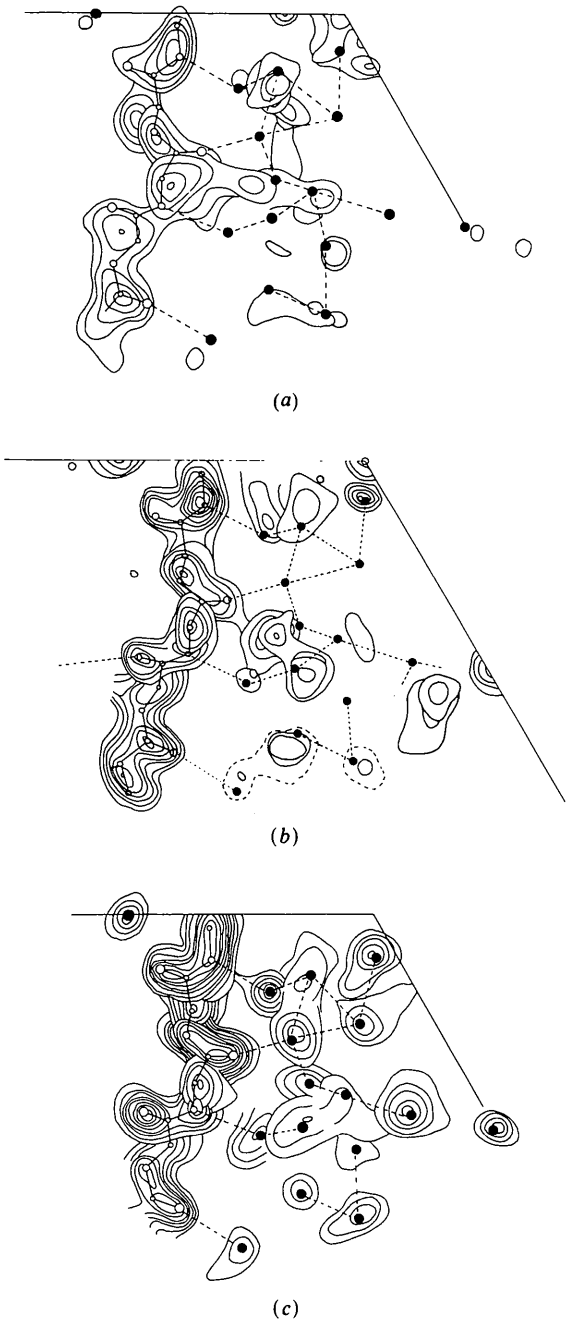

Fig. 7. Comparison of three electron density views for residues B3 to B7 of molecule 1 and the surrounding solvent structure. The maps were constructed using phases determined by: (*a*) multiple isomorphous replacement method (1·9 Å resolution); (*b*) phase refinement and extension from 1·9 to 1·5 Å resolution; (*c*) refinement of atomic positions by least-squares minimization and Fourier methods (1·5 Å resolution). These sections are from a difference Fourier map calculated with observed moduli and phases corresponding to the refined structure in which $\frac{1}{8}$ of the unit-cell volume is left out. The protein features are clearly present in all maps but most resolved in (*b*) and (*c*). Few water molecules, which are mostly ill defined, can be seen in (*a*). In (*b*), one water molecule which is coordinated to zinc on the threefold axis and is well defined does appear. The overall character of the water and peptide density is well defined in (*c*).

interesting to note that the maps calculated after phase extension reveal the structure of well defined water molecules clearly and usually in correct positions. To clarify this point, Fig. 9 indicates as a function of the final temperature factor the number of water molecules present in the maps compared

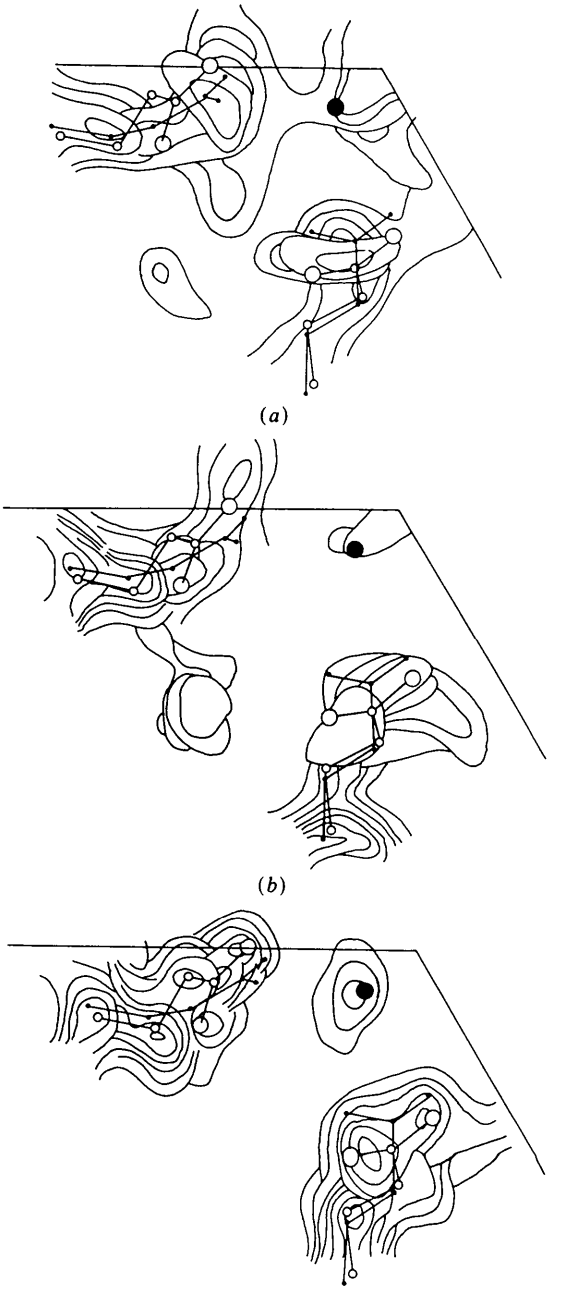

Fig. 8. Comparison of three electron density maps for residues B13 of molecules 1 and 2. ○ Refined atomic positions (1·5 Å); ● initial atomic positions (1·9 Å). Maps are constructed by using phases determined by: (*a*) multiple isomorphous replacement method (1·9 Å spacing); (*b*) phase refinement and extension from 1·9 to 1·5 Å spacing; (*c*) refined atomic positions and thermal parameters (same conditions as for Fig. 7*c*).

with those introduced in the atomic-coordinate refinement.

The top of Fig. 7(*b*) gives a picture of an electron density region close to a Zn atom where many water molecules are present. Although the quality of this area is not as good as in the final electron density maps (Fig. 7*c*), it is clear that some water molecules come out more precisely than in the initial maps (Fig. 7*a*). For example, near the Zn threefold site, an area of diffuse electron density in the isomorphous map, a water molecule appears clearly without any associated spurious peaks.

Conclusion

These calculations have demonstrated that experimental phases can be improved and extended usefully by the maximum determinant procedure. The method benefits greatly from the reasonably good and extensive set of experimental phases and a high-resolution set of native data. The improvement in the phase angles from the isomorphous values is substantial and extends to the limit of the phase data. Beyond 1·9 Å, the limit of phased data, the errors in the extended phase (for terms $E > 1$) are remarkably small (38°). Our conclusions are:

1. The experiments of this study and Sayre (1972) with rubredoxin (both with 1·5 Å data) show that with present procedures the ability of direct-method techniques to improve phases is conditioned by the quality and resolution of the phases available. At the other extreme, phase determination with a limited series based on solvent modelling has proved possible and has helped to produce a satisfactory set of low-order phases (Podjany, Schevitz & Sigler, 1981; Schevitz *et al.*, 1981). These results demonstrate that the structural information in the high-angle data is not lost in spite of their inaccuracy; with sufficiently good starting phases, it can be reconstructed. Thus the challenges of improving poor phases or extending phases from low resolution can be faced with optimism.

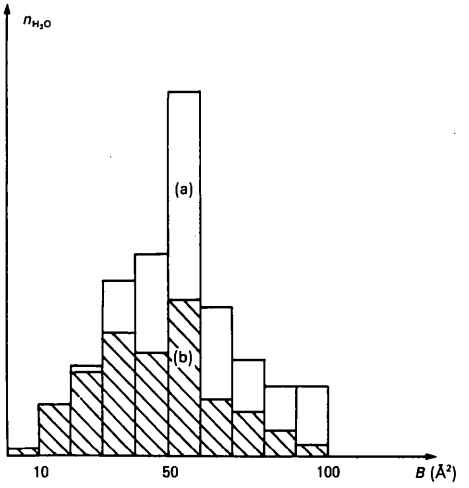

Fig. 9. Distribution of the number of water molecules as a function of associated temperature factors (*a*) introduced in the atomic-coordinate refinement; (*b*) placed in the maps constructed after phase refinement and extension from 1·9 to 1·5 Å spacing.

2. The well behaved atoms with $B < B_{av}$ generally are better defined and more accurately positioned. The poorly ordered atoms such as protein side chain and solvent atoms with $B > B_{av}$ tend to have much reduced or no electron density. It was not possible to assign atomic positions for these atoms with any confidence. This effect in some ways simplifies the interpretation of the map and subsequent refinement by limiting the derived coordinates to those which have low B values.

3. The effectiveness of this determinantal procedure and that of least-squares refinement procedures are similar but cannot be compared directly. The 1·9 Å isomorphous map was generally very good and would have been readily interpreted for the most part but of course the structure had already been defined at 2·8 Å resolution. The extension and refinement of the 1·9 Å isomorphous phases, on the other hand, took relatively little time and led to a map from which more accurate coordinates could be determined and where poorly defined structure was recognizable. The atomic coordinates derived from the 1·9 Å resolution map contained errors while those derived from the 1·5 Å solution phase extended map were incomplete. Both sets nonetheless refined smoothly from 48 to 33% and 40 to 27%, respectively, in the early stages. Most importantly, they produced difference Fourier maps in which the poorly defined structure could be sensibly analysed.

4. The need at present for reasonably good phases and extensive diffraction data means that the principal advantage of the determinantal method (that atomic parameters are not needed to extend and improve the phases) is most useful for proteins where sequence information is absent or incomplete.

This work was carried out with the support of the CNRS, the British Diabetic Association, the Medical Research Council and the Kroc Foundation.

APPENDIX I

Variation of the occupancy factor as a function of r and R

If the basic elements are chosen randomly inside a sphere of radius r, the reciprocal vector **h** associated with one basic element can be considered as a random variable (r.v.) and the corresponding density of probability (d.p.) is given by

$$f(\mathbf{h}) = \begin{cases} 0 & \text{if } |h| > r \\ 1/V & \text{if } |h| \leq r \end{cases}$$

with $V=(4/3)\pi r^3$, volume of the sphere. Here, the r.v. is considered as a continuous variable. This is justified because of the large number of reflexions inside the sphere.

Now, let us consider the vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ associated with two basic elements and the r.v.:

$$\mathbf{h}_{12}=\mathbf{h}_1-\mathbf{h}_2.$$

The variables $\mathbf{h}_1$ and $\mathbf{h}_2$ are independent and identical to $\mathbf{h}$, hence the d.p. of $\mathbf{h}_{12}$ is given by

$$g(\mathbf{h}_{12})=\int_{R^3} f(\mathbf{h}_{12}+\mathbf{h}_2)f(\mathbf{h}_2)\,\mathrm{d}^3\mathbf{h}_2$$

$$g(\mathbf{h}_{12})=\begin{cases}0 & \text{if } |\mathbf{h}_{12}|>2r\\ (v(\mathbf{h}_{12})/v^2) & \text{if } |\mathbf{h}_{12}|<2r\end{cases},$$

where $v(\mathbf{h}_{12})$ is the volume of the common part of two spheres of radius r with centres O and O' such that

$$OO'=\mathbf{h}_{12}.$$

The $v(\mathbf{h}_{12})$ is determined by

$$v(\mathbf{h}_{12})=2\int_{D/2}^{r}\pi(r^2-z^2)\,\mathrm{d}z$$
$$=(2\pi/3)(r-D/2)^2(2r+D/2)$$

where $D=|\mathbf{h}_{12}|$.

The d.p. of $g(\mathbf{h}_{12})$ can be expressed by

$$g(\mathbf{h}_{12})=\begin{cases}0 & \text{if } D>2r\\ (2\pi r^3/3V^2)(1-D/2r)^2(2+D/2r) & \text{if } D\le 2r\end{cases}$$

If we make use of the r.v. $u=\mathbf{h}_{12}/2r$ with modulus u,

$$g(\mathbf{h}_{12})=\begin{cases}0 & \text{if } u>1\\ (1/2V)(1-u^2)(2u) & \text{if } u\le 1\end{cases},$$

and take into account the spherical symmetry of $g(h_{12})$, the d.p. of the r.v. D is given by

$$f(D)=g(\mathbf{h}_{12})4\pi D^2$$

or

$$f(D)=\begin{cases}0 & \text{if } u>1\\ 6u^2(1-u^2)(2+u)/r & \text{if } u<1\end{cases},$$

which leads directly to the d.p. $G(u)$:

$$G(u)=\begin{cases}0 & \text{if } u>1\\ 12u^2(1-u^2)(2+u) & \text{if } u<1\end{cases}.$$

A representation of this function is given in Fig. 1 (§ II). As we expect, it is equal to 1 if $r<\frac{1}{2}$ and then decreases to zero when r increases.

APPENDIX II

To assess the quality of the matrix $\|U\|_m$ a general criterion is that the value of the determinant should be as low as possible. This can be justified by the following (de Rango, 1969).

The regression formula is derived from the probability law of m structure factors ($E_1=E_{\mathbf{L}},\ldots,E_p=E_{\mathbf{L}-\mathbf{H}_p},\ldots,E_m=E_{\mathbf{L}-\mathbf{H}_m}$), which is a Gaussian multidimensional law defined by a covariance matrix $\|U\|_m$ whose determinant is D_m.

The *a priori* knowledge of the autocorrelation will diminish the entropy of the distribution which can be estimated by the Shannon information theorem (Shannon & Weaver, 1949). In the case of m-dimensional continuous distribution, the entropy is defined by

$$I(E_1,\ldots,E_p,\ldots,E_m)$$
$$=-\int\ldots\int\ldots\int p(E_1,\ldots,E_p,\ldots,E_m)$$
$$\times\log_e p(E_1,\ldots,E_p,\ldots,E_m)\,\mathrm{d}E_1\ldots\mathrm{d}E_p\ldots\mathrm{d}E_m.$$

When the distribution is a Laplace Gauss law it has been shown (Renyi, 1966) that

$$I(E_1,\ldots,E_p,\ldots,E_m)$$
$$=\log_e(2\pi e)^{n/2}D_m^{1/2}\quad\text{for real variables}$$

and

$$I(E_1,\ldots,E_p,\ldots,E_m)$$
$$=\log_e(\pi e)^m D_m\quad\text{for complex variables.}$$

If we consider the random variables $E_1,\ldots,E_p,\ldots,E_m$ as independent variables, the value of D_m is equal to 1 and the entropy is equal to $m\log_e(\pi e)$. But, when we define the variables $E_1,\ldots,E_p,\ldots,E_m$ by imposing the autocorrelation (*i.e.* all the elements $U_{\mathbf{H}_p-\mathbf{H}_q}$ of the matrix $\|U\|_m$ are fixed and known), the entropy of the distribution decreases since the quantity $I_G=\log D_m$ is always negative ($D_m<1$).

So, the lower the value of D_m, the higher the gain of information (decrease of entropy) related to the knowledge of $\|U\|_m$. The I_G may constitute a quantitative criterion of selection for the Karle–Hauptman determinant to be used in a phase refinement and extension procedure.

References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.
BRICOGNE, G. (1976). *Acta Cryst.* A**32**, 832-847.
CASTELLANO, E., PODJARNY, A. D. & NAVAZA, J. (1973). *Acta Cryst.* A**29**, 609-615.
COLLINS, D. M. (1975). *Acta Cryst.* A**31**, 388-389.
COLLINS, D. M. (1978). *Acta Cryst.* A**34**, 533-541.
COLLINS, D. M. (1982). *Nature (London)*, **298**, 49-51.
DODSON, E. J., ISAACS, N. W. & ROLLETT, J. S. (1976). *Acta Cryst.* A**32**, 311-315.
DODSON, G. G., DODSON, E. J., HODGKIN, D. C., ISAACS, N. W. & VIJAYAN, M. (1975). Unpublished results.
DODSON, G. G., DODSON, E. J., HODGKIN, D. C. & REYNOLDS, C. D. (1979). *Can. J. Biochem.* **68**, 469-479.
ELLER, G. VON (1962). *Acta Cryst.* **15**, 590-595.
ELLER, G. VON (1964). *C.R. Acad. Sci.* **258**, 5418-5419.
GASSMANN, J. (1966). Thesis. Technical Univ. Munich.
GASSMANN, J. (1976). *Acta Cryst.* A**32**, 274-280.

GOEDKOOP, J. A. (1952). *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, edited by R. PEPINSKY. Pennsylvania State College, Univ. Press.
HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136-143.
HENDRICKSON, W. A., LOVE, W. E. & KARLE, J. (1973). *J. Mol. Biol.* **74**, 331-361.
KARLE, J. & HAUPTMAN, H. (1950). *Acta Cryst.* **3**, 181-187.
KNOSSOW, M., DE RANGO, C., MAUGUEN, Y., SARRAZIN, M. & TSOUCARIS, G. (1977). *Acta Cryst.* A**33**, 119-125.
LAJZEROWICZ, J. & LAJZEROWICZ, J. (1966). *Acta Cryst.* **21**, 8-12.
MAUGUEN, Y. (1979). Thesis, Paris.
NAVAZA, J. & SILVA, A. M. (1979). *Acta Cryst.* A**35**, 266-275.
PODJARNY, A. D., SCHEVITZ, R. W. & SIGLER, P. (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**, 281-294.
RAGHAVAN, N. V. & TULINSKY, A. (1979). *Acta Cryst.* B**35**, 1776-1785.
RANGO, C. DE (1969). Thesis, Paris.
RANGO, C. DE, MAUGUEN, Y. & TSOUCARIS, G. (1975). *Acta Cryst.* A**31**, 227-233.
RANGO, C. DE, MAUGUEN, Y. & TSOUCARIS, G. (1978). *Acta Cryst.* A**34**, S47.
RANGO, C. DE, MAUGUEN, Y., TSOUCARIS, G., CUTFIELD, J., DODSON, G. G. & ISAACS, N. W. (1975). *Acta Cryst.* A**31**, S21.
RANGO, C. DE, MAUGUEN, Y., TSOUCARIS, G., DODSON, G. G., DODSON, E. J. & TAYLOR, D. J. (1979). *J. Chim. Phys.* **76**, No. 9, 811-812.
RANGO, C. DE, TSOUCARIS, G. & ZELWER, C. (1974). *Acta Cryst.* A**30**, 342-353.
RENYI, A. (1966). *Calcul des Probabilités avec un Appendice sur la Théorie de l'Information.* Paris: Dunod.
ROSSMANN, M. G. & BLOW, D. M. (1961). *Acta Cryst.* **14**, 641-647.
SAYRE, D. (1972). *Acta Cryst.* A**28**, 210-212.
SAYRE, D. (1974). *Acta Cryst.* A**30**, 180-184.
SCHEVITZ, R. W., PODJARNY, A. D., ZWICK, M., HUGHES, J. J. & SIGLER, P. (1981). *Acta Cryst.* A**37**, 669-677.
SHANNON, G. E. & WEAVER, W. (1949). *The Mathematical Theory of Communication.* Univ. of Illinois Press.
SIM, G. A. (1960). *Acta Cryst.* **13**, 511-512.
TAYLOR, D. J., WOOLFSON, M. M. & MAIN, P. (1978). *Acta Cryst.* A**34**, 870-883.
TSOUCARIS, G. (1970). *Acta Cryst.* A**26**, 492-499.
TSOUCARIS, G. (1981). In *Theory and Practise of Direct Methods in Crystallography*, edited by M. F. C. LADD & R. A. PALMER, pp. 282-360. New York: Plenum Press.
WOOLFSON, M. M. (1956). *Acta Cryst.* **9**, 804-810.

Acta Cryst. (1985). A**41**, 17-25

X-ray Diffraction Lenses with Spherical Focusing

BY V. I. KUSHNIR

Institute of Solid State Physics, USSR Academy of Sciences, 142432 *Chernogolovka, Moscow Region, USSR*

V. M. KAGANER

Institute of Crystallography, USSR Academy of Sciences, 59 *Leninskii Prospekt,* 117333 *Moscow, USSR*

AND E. V. SUVOROV

Institute of Solid State Physics, USSR Academy of Sciences, 142432 *Chernogolovka, Moscow Region, USSR*

(*Received* 4 *January* 1984; *accepted* 20 *June* 1984)

Abstract

X-ray spherical-wave focusing in multibeam dynamical diffraction by a biaxially bent single crystal has been considered. In contrast to cylindrical lenses already studied in the two-beam case, which presented a line focus, here wave packets focusing in two directions into a single point are dealt with. The conditions for focusing of the trajectories of the X-ray Bloch waves are established and the algorithm for the determination of the parameters of corresponding X-ray optical systems is described. Possible sets of parameters are calculated. The X-ray wave field distribution in a crystal is simulated numerically. The calculated topographs confirm the existence of the focusing effect.

1. Introduction

Optics for focusing X-rays and neutrons have been developed in two main directions - Fresnel zone plates for soft radiation (~40 Å) (Kirz, 1974; Kearney, Klein, Opat & Gähler, 1980) and diffraction lenses based on single crystals for hard radiation (~1 Å). The operation of such lenses is based on the effect of dynamical focusing of X-rays (or neutrons), which consists of the compression of coherent wave packets during dynamical scattering by single crystals. Until recently both theorists and experimenters have studied this effect mainly for two-beam diffraction (see references in the paper by Kushnir & Suvorov, 1982). In this case wave packets can be compressed by a crystal only in one direction determined by the

0108-7673/85/010017-09$01.50