acter it is conceivable that the correction term might still become significant if the diffracting region is long. It is difficult to say exactly when the correction term in equation (26) might become important, but it could be a useful tool in numerical calculations to check the validity of ignoring $\partial^2\mathbf{D}(x,z)/\partial x^2$ in the second-order equations. By simultaneously evaluating the first correction term in equation (26) while performing the numerical integration of the hyperbolic system, one can tell immediately when this term becomes a significant correction to $\mathbf{D}(x,z)$.

We see that a good working result is the following: Let $\mathbf{D}^{\mathrm{I}}(x,t)$ be a solution (*e.g.* obtained numerically) to the first-order equation for a crystal of thickness t. Then, we may have confidence in this solution if

$$\left|\int_0^t (t-z')\mathbf{C}\frac{\partial^2}{\partial x^2}[\mathbf{A}(x^{\mathrm{ret}\cdot},z')\mathbf{D}^{\mathrm{I}}(x^{\mathrm{ret}\cdot},z')]\mathrm{d}z'\right| \ll|\mathbf{D}^{\mathrm{I}}(x,t)| \quad (27)$$

for a range of x, and a number of components of $\mathbf{D}^{\mathrm{I}}(x,t)$ for which $\mathbf{D}^{\mathrm{I}}(x,t)$ is numerically significant.

5. Conclusions

By converting the dynamical equations of high-energy-electron diffraction from an imperfect crystal into integral equations we have been able to (i) construct an iterative solution for the Fourier coefficients of the electron wave function for both the first and second-order equations, and (ii) compare these solutions in such a fashion as to obtain an explicit correction term, which is a measure of the validity of ignoring second-order partial derivatives in the dynamical equations.

In concluding this paper, we would like to stress the complementary aspects of the differential equation approach and the integral approach to high-energy-electron diffraction theory. The differential equations have received the most attention to date, undoubtedly owing to the ease by which they can be numerically implemented. On the other hand, the integral equations that we have developed in this paper have definite theoretical advantages. For instance, with regard to the approximation investigated in this paper, it is difficult to look at the differential equation (7) and assess the importance of the $-i\nabla^2 d_g(x,z)/2k_{gz}$ term. Even though the coefficient, $-i/2k_{gz}$, of this term is small, in some sense, we intuitively realize that it must be the specific crystal potential that decides the matter. In other words, there may be cumulative effects that make this term important. These cumulative effects are explicitly displayed in the integral of equation (27) thereby confirming our intuitive feelings.

References

COPSON, E. T. (1965). *Asymptotic Expansions*. Cambridge Univ. Press.

HOWIE, A. & BASINSKI, Z. S. (1968). *Phil. Mag.* **17**, 1039–1063.

KURIYAMA, M. (1972). *Acta Cryst.* A**28**, 588–593.

LEWIS, A. L. & VILLAGRANA, R. E. (1972). *Phys. Rev.* B**6**, 4382–4392.

TAKAGI, S. (1962). *Acta Cryst.* **15**, 1311–1312.

TAKAGI, S. (1969). *J. Phys. Soc. Japan*, **26**, 1239–1253.

Acta Cryst. (1975). A**31**, 227

Use of High-Order Probability Laws in Phase Refinement and Extension of Protein Structures

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

Laboratarie de Physique, Tour B, Centre Universitaire Pharmaceutique, 92290 *Chatenay Malabry, France*

(*Received* 21 *November* 1973; *accepted* 16 *July* 1974)

High-order covariance matrices are used to show that the maximal determinant rule and the regression equation can be applied successfully to the phase refinement and extension of protein structures. With structure factors calculated from the atomic model of insulin, the use of an order-400 covariance matrix leads to the structure phases with an average error $\Delta\Phi$ of 15°. The method has also been applied to actual data of insulin for phase refinement and for phase extension from 2·8 to 2 Å.

The investigation of the probability law for one structure factor included in a Karle–Hauptman determinant, connected with the regression-plane equation, has been developed recently (de Rango, 1969; Tsoucaris, 1970; de Rango, Tsoucaris & Zelwer, 1974). Here the possibility of applying the regression equation in phase determination of protein structures is investigated in two important cases:

– refinement of the phases approximately known from the isomorphous-replacement method,

– phase extension: determination of new structure-factor phases from approximately known data.

We have already shown that the maximal determinant rule and the regression equation, using high-order covariance matrices, can be applied sucessfully to the phase determination of protein structures (de

Rango & Mauguen, 1972). This has been performed with an order-150 covariance matrix for the structures of myoglobin (Kendrew, Dickerson, Stranberg, Davies, Phillips & Shore, 1960) and insulin (Blundell, Cutfield, Dodson, Dodson, Hodgkin, Mercola & Vijayan, 1971).

In the present paper, we report the essentials of these results and also show that the use of high-order matrices leads to more accurate phases when the order increases from 150 to 400 (de Rango, Tsoucaris & Mauguen, 1973).

Theoretical background

We recall first the essential results of multivariate probability theory. Let us consider a set of m normalized structure factors assumed to be random variables (*i.e.* unknown):

$$E_q = E_{\mathbf{L}-\mathbf{H}_q} \quad q=1 \ldots m \qquad \begin{matrix}\mathbf{L}\text{: variable vector} \\ \mathbf{H}_q\text{: fixed vector.}\end{matrix}$$

The $m \times m$ table of the correlation coefficients U_{pq} between these m structure factors form the *covariance matrix*, the determinant of which is a Karle–Hauptman determinant D_m [see below equation (6)].

These correlation coefficients are given by the Sayre–Hughes equation, as can be seen below:

$$\langle E_{\mathbf{L}-\mathbf{H}_p} E^*_{\mathbf{L}-\mathbf{H}_q} \rangle_{\mathbf{L}} = U_{qp} = U_{\mathbf{H}_q-\mathbf{H}_p} \tag{1}$$

$$\det (U_{qp}) = D_m \geq 0 \,. \tag{1a}$$

The elements of the covariance matrix are unitary structure factors ($U_{000}=1$).

It has been shown that the multivariate probability density $p(E_1 \ldots E_m)$ of the $E_q = E_{\mathbf{L}-\mathbf{H}_q}$ set ($q=1 \ldots m$) is an m-dimensional Laplace–Gauss law, as follows† (in the non-centrosymmetric case):

If we write

$$Q_m = \sum_{q=1}^{m} D_{qq} |E_q|^2 + \sum_{\substack{p=1 \\ p \neq q}}^{m} \sum_{q=1}^{m} D_{pq} E_p E_q^* \tag{2}$$

where D_{pq} is an element of the inverse matrix of U_{qp}, then

$$p(E_1 \ldots E_m) = \text{constant} \,.\, \exp(-Q_m) \,. \tag{3}$$

Clearly, the maximum probability occurs for

$$Q_m = \text{minimum} \,. \tag{4}$$

If we denote by $\bar{E}_q$, $q=1 \ldots m$, the value of E_q for which Q_m is minimum, we can write:

$$\left[\frac{\partial Q_m}{\partial E_q}\right]_{\bar{E}_q} = 0 \,. \tag{4a}$$

If E_q is real, this is obvious; if E_q is complex, we make the derivation separately for the real and imaginary parts.

† The notation is explained in detail in a previous paper (de Rango, Tsoucaris & Zelwer, 1974). Here, we will consider only the non-centrosymmetric case.

By substituting equation (2) in equation (4*a*), we obtain the regression equation:

$$\bar{E}_q = |\bar{E}_q| \exp(i\bar{\varphi}_q) = -\frac{1}{D_{qq}} \sum_{\substack{p=1 \\ p \neq q}}^{m} D_{pq} E_p \,. \tag{5}$$

The above results can be connected with determinants and expressed alternatively in the form of the *maximum determinant rule*, which is strictly equivalent to equation (4) or (4*a*):

$$(\Delta_{m+1})_{\substack{\text{most probable} \\ \text{values of } E_q\text{'s}}} = \text{maximum} \tag{6}$$

where Δ_{m+1} is the following determinant (the dots are drawn only to emphasize the special role of the last row and column which contain the random variables, *i.e.* the unknown structure factors):

$$\Delta_{m+1} = \frac{1}{N} \begin{vmatrix} 1 & \ldots & U_{-\mathbf{H}_p} & \ldots & U_{-\mathbf{H}_q} & \ldots & E_{-\mathbf{L}} \\ \vdots & & \vdots & & \vdots & & \vdots \\ U_{\mathbf{H}_p} & \ldots & 1 & \ldots & U_{\mathbf{H}_p-\mathbf{H}_q} & \ldots & E_{-\mathbf{L}+\mathbf{H}_p} \\ \vdots & & \vdots & & \vdots & & \vdots \\ U_{\mathbf{H}_q} & \ldots & U_{\mathbf{H}_q-\mathbf{H}_p} & \ldots & 1 & \ldots & E_{-\mathbf{L}+\mathbf{H}_q} \\ \vdots & & \vdots & & \vdots & & \vdots \\ E_{\mathbf{L}} & \ldots & E_{\mathbf{L}-\mathbf{H}_p} & \ldots & E_{\mathbf{L}-\mathbf{H}_q} & \ldots & N \end{vmatrix} .$$

If in Δ_{m+1}, in addition to the known U_{qp}'s, all elements of the last row are known except one, say E_q, then its expected value $\bar{E}_q$ is given precisely by the regression equation (5). In this work only this equation has been used; however, an algorithm to minimize Q_m by using an eigenvector procedure has been derived by Sarrazin (1970).

Next, probability theory shows that $E_q = A_q + iB_q$ is distributed in the complex plane, according to a Gaussian law:

$$p(E_q) = p(A_q, B_q) = \frac{1}{\pi\sigma_q^2} \exp\left(-\frac{|E_q - \bar{E}_q|^2}{\sigma_q^2}\right) \tag{7}$$

where $\bar{E}_q$ is the expected value [equation (5)] and σ_q^2 is the variance given by:

$$\sigma_q^2 = \frac{1}{D_{qq}} < 1 \,. \tag{7a}$$

If we assume now that $|E_q|$ is known, the distribution of the phase difference ε_q between φ_q (exact phase) and $\bar{\varphi}_q$ (expected phase) is given by:

$$p(\varepsilon_q/|E_q|) = \frac{\exp(A_q \cos \varepsilon_q)}{\pi I_0(A_q)} \,; \quad \varepsilon_q = \varphi_q - \bar{\varphi}_q \tag{8}$$

$$A_q = 2|E_q \bar{E}_q D_{qq}| \,. \tag{8a}$$

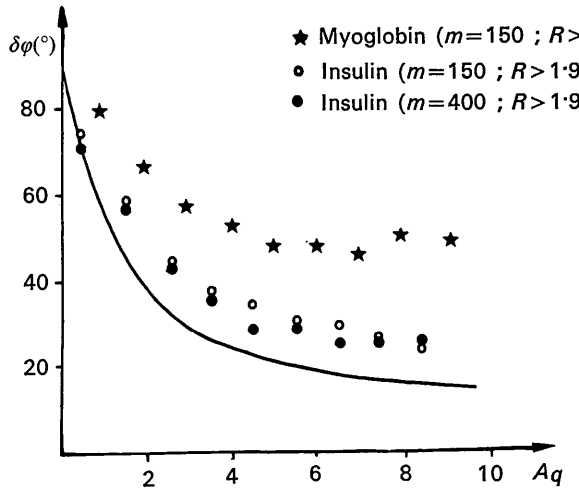

Fig. 1. Variation of $\delta\varphi$ as a function of A_q for myoglobin and insulin. The curve indicates the theoretical distribution given by equation (14).

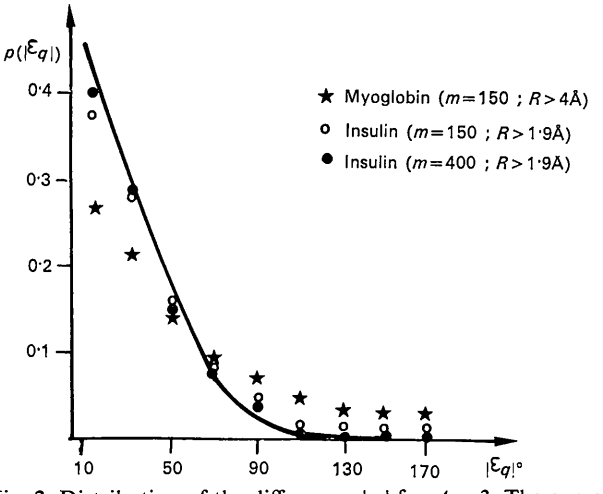

Fig. 2. Distribution of the differences $|\varepsilon_q|$ for $A_q=3$. The curve indicates the theoretical distribution given by equation (8).

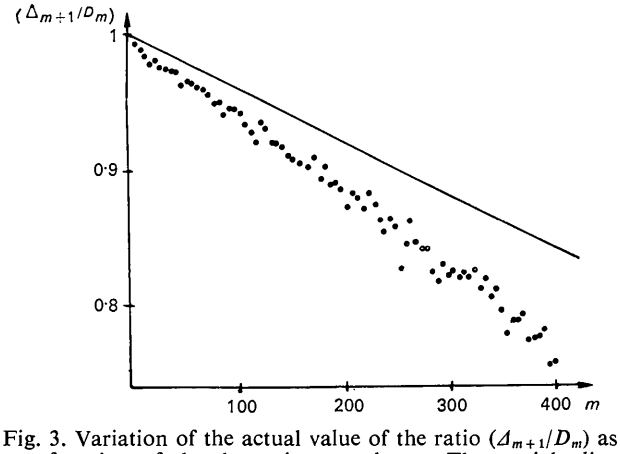

Fig. 3. Variation of the actual value of the ratio (Δ_{m+1}/D_m) as a function of the determinant order m. The straight line indicates the theoretical mean value $(1-m/N)$.

Phase-determination process

The principle resides in the fact that, in the above theory, $\mathbf{L}$ is a random vector. When $\mathbf{L}$ sweeps out reciprocal space (the $\mathbf{H}_p$'s being fixed), this generates a family of $\Delta_{m+1}(\mathbf{L})$ determinants having the same principle minor D_m. Also, one can construct several determinants D_m, and generate in the same way several families of $\Delta_{m+1}(\mathbf{L})$.

In all cases, the same structure factor (or a symmetry related to it) say $E_{\mathbf{L}-\mathbf{H}_q}$, may appear several times in the last row of different determinants, labelled by $1,\ldots i,j\ldots$ For instance, the reflexion $\mathbf{H}=\mathbf{L}_j-\mathbf{H}_q$ may appear in the jth determinant $\Delta^{(j)}_{m+1}$ *and* in the ith determinant $\Delta^{(i)}_{m+1}$, provided that:

$$\mathbf{H}=\mathbf{L}_i-\mathbf{H}_p=\mathbf{L}_j-\mathbf{H}_q=\ldots . \tag{9}$$

For each of these determinants, a different expected value is determined by equation (5) for the same reflexion $E_{\mathbf{H}}=|E_{\mathbf{H}}|\exp(i\varphi_{\mathbf{H}})$.

By taking into account equation (9), these determinations will be denoted as follows:

$$\bar{E}_{\mathbf{H}}^{(j)}=\bar{E}_{\mathbf{L}j-\mathbf{H}_q} \text{ from } \Delta^{(j)}_{m+1}$$
$$\bar{E}_{\mathbf{H}}^{(i)}=\bar{E}_{\mathbf{L}i-\mathbf{H}_p} \text{ from } \Delta^{(i)}_{m+1} \quad etc.\ldots .$$

A better approximation to $E_{\mathbf{H}}$ will be obtained by averaging over all such determinations

$$E_{\mathbf{H}}\simeq\langle\bar{E}_{\mathbf{H}}^{(j)}\rangle_j . \tag{10}$$

Equation (10) is strictly valid, of course, only if all determinations are independent. After introducing a weighting factor:

$$1/\sigma_q^{2(j)}=D_{qq}^{(j)}$$

we have

$$E_{\mathbf{H}}\simeq\langle D_{qq}^{(j)}\bar{E}_{\mathbf{H}}^{(j)}\rangle_j . \tag{10a}$$

Finally, the phase Φ_H of the right-hand side of (10*a*) is

$$\tan\Phi_H=\frac{\sum_j D_{qq}^{(j)}|\bar{E}_{\mathbf{H}}^{(j)}|\sin\bar{\varphi}_{\mathbf{H}}^{(j)}}{\sum_j D_{qq}^{(j)}|\bar{E}_{\mathbf{H}}^{(j)}|\cos\bar{\varphi}_{\mathbf{H}}^{(j)}} \tag{10b}$$

with

$$\bar{E}_{\mathbf{H}}^{(j)}=|\bar{E}_{\mathbf{H}}|^{(j)}\exp(i\overline{\varphi_{\mathbf{H}}}^{(j)}) \tag{10c}$$

and can be associated with the probability law

$$p(\varphi_{\mathbf{H}}-\Phi_{\mathbf{H}})=\frac{1}{\pi I_0(A_{\mathbf{H}})}\exp[A_{\mathbf{H}}\cos(\varphi_{\mathbf{H}}-\Phi_{\mathbf{H}})] \tag{11}$$

where $\varphi_{\mathbf{H}}$ is the exact phase, I_0 denotes a Bessel function and $A_{\mathbf{H}}$ is given by

$$A_{\mathbf{H}}^2=[\sum_j A_{\mathbf{H}}^{(j)}\cos\overline{\varphi_{\mathbf{H}}}^{(j)}]^2+[\sum_j A_{\mathbf{H}}^{(j)}\sin\overline{\varphi_{\mathbf{H}}}^{(j)}]^2. \tag{11a}$$

In this last equation $A_{\mathbf{H}}^{(j)}$ indicates the value of A_q, as given by equation (8*a*) associated with the occurence

of reflexion $\mathbf{H}=\mathbf{L}_j-\mathbf{H}_q$ at the **q**th column of the jth determinant $\Delta_{m+1}^{(j)}$.

The practical calculation of the above formulae could be speeded up by using Goedkoop determinants. However, it has been shown that the phases obtained by the maximal determinant rule are identical whether one uses Karle–Hauptman determinants containing all equivalent reflexions, or Goedkoop determinants (Mauguen, de Rango & Tsoucaris, 1973).

The relation (5) should be considered as a statistical relation and not as a strict equality (for $m<N$). This implies that there is no rational dependence between the coordinates. In other words, the distribution of the structure invariants included in the determinant should be similar to the distribution given by Cochran (1955). Now, the structure-invariant distribution corresponding to low-index reflexions deviates from this distribution, because too many high-modulus structure invariants occur with phase too far from zero ('abnormal invariants'). Also, it is important to emphasize that, in general, the Laplace–Gauss m-dimensional law should not be strictly valid for a resolution lower than 3 Å.

Some important aspects of the efficiency of the above equations were considered in a previous paper (de Rango, Tsoucaris & Zelwer, 1974). From this discussion and the above remarks it appears that, in the phase determination of protein structures, the regression equation (5) should be applied to D_m determinants for which:

– the value should be as small as possible,

– the occurence of 'abnormal invariants' should be avoided as much as possible.

Analysis of these results – application to model structures

In order to test the validity of the theoretical formulae we first use normalized structure factors calculated from the atomic models of myoglobin and insulin.

A measure of the overall efficiency of the method, for a given set of reflexions, may be obtained from the average absolute value of the differences between the exact value φ_H and the computed value $\Phi_{\mathbf{H}}$ [equation (10*b*)]:

$$\Delta\Phi=\langle|\varphi_{\mathbf{H}}-\Phi_{\mathbf{H}}|\rangle_{\mathbf{H}} \qquad (12)$$

when **H** belongs to a given set.

A measure of the precision associated with the determination of each contributor to tan $\Phi_{\mathbf{H}}$ in equation (10*b*) is also evaluated, as follows: by recalling that $\varphi_{\mathbf{H}}^{(j)}$ is in fact the value of $\bar{\varphi}_q=\bar{\varphi}_{\mathbf{L}-\mathbf{H}_q}$ obtained from the jth determinant, we define:

$$\delta\varphi=\langle|\varepsilon_q|\rangle_{q,\mathbf{L}}=\langle|\varphi_{\mathbf{L}-\mathbf{H}_q}-\bar{\varphi}_{\mathbf{L}-\mathbf{H}_q}|\rangle_{q,\mathbf{L}} \qquad (13)$$

where $\varphi_q=\varphi_{\mathbf{L}-\mathbf{H}_q}$ is the exact value and $\bar{\varphi}_q=\bar{\varphi}_{\mathbf{L}-\mathbf{H}_q}$ is a single determination obtained by equation (5). Here, also, the variable indices q and **L** may belong to a given set which will be indicated at each use of (13).

Influence of the resolution

For the 4 Å resolution, order-151 determinants have been built up by choosing the first-row elements (called basic elements) between the $|E|$'s of high modulus included in the polar sphere $R>8$ Å. In this way, all elements of D_m are included in the polar sphere $R>4$ Å. For myoglobin, the phases of 322 structure factors (with $|E|>1$) have been calculated by using 150 order-151 determinants Δ_{m+1}. The overall average error $\Delta\Phi$ is 45°.

Fig. 1 shows the variation of $\delta\varphi$ as a function of A_q, defined by equation (8*a*). This means that the average of equation (13) has been performed separately for those terms for which A_q has values in the intervals

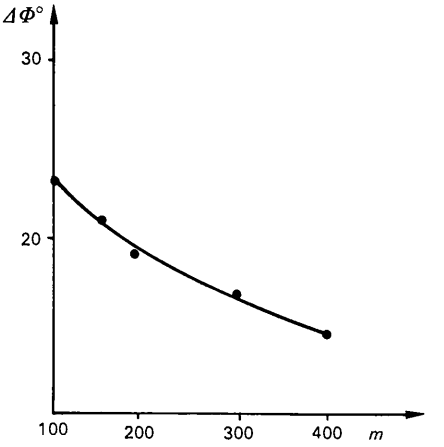

Fig. 4. Variation of the overall average value $\Delta\Phi$ as a function of the determinant order. (1·9 Å resolution).

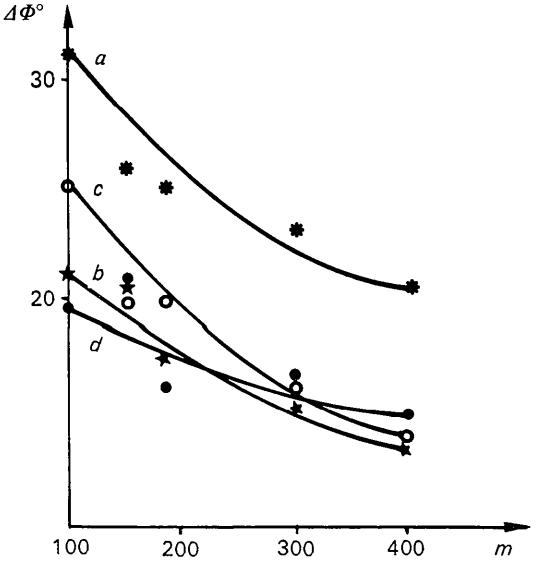

Fig. 5. Variation of $\Delta\Phi$ as a function of the determinant order and of the resolution R. Curve *a*: $R>6$ Å; Curve *b*: 6 Å $>R>4$ Å; Curve *c*: 4 Å $>R>2{\cdot}5$ Å; Curve *d*: 2·5 Å $>R>$ 1·9 Å.

indicated in Fig. 1. The corresponding theoretical curve is given by:

$$(\delta\varphi)_{\text{theoretical}} = \frac{1}{\pi}\int_0^{\pi} \varepsilon p(\varepsilon)\mathrm{d}\varepsilon \qquad (14)$$

where $p(\varepsilon)$ is the probability density given by equation (8).

The obtained plot is far from the theoretical curve, especially for the higher values of A_q. Moreover, the histogram of $|\varepsilon_q|$ for a given A_q is not in good agreement with the theoretical distribution given by equation (8) even for medium values of A_q, as shown in Fig. 2. Similar results have been obtained for insulin.

For the 1·9 Å resolution, the calculation has been performed only with insulin. An appreciable improvement has been obtained in these results. The order of D_m is still 150, but the low-index reflexions which involve 'abnormal' invariants, are left out. Symmetric reflexions are forbidden as basic elements, which are chosen from the $|E|$'s of high modulus included in the polar sphere $R>3{\cdot}8$ Å. By using 800 determinants Δ_{m+1}, the phases of 1264 structure factors with modulus higher than one have been calculated with an average error $\Delta\Phi$ of 21°; these results† are summarized in Table 1. The variation of $\delta\varphi$ [equation (13)] as a function of A_q is now closer to the theoretical curve

† The use of 800 order-151 determinants allows the determination of a thousand structure-factor phases in less than 5 min with a computer such as the IBM 370/165.

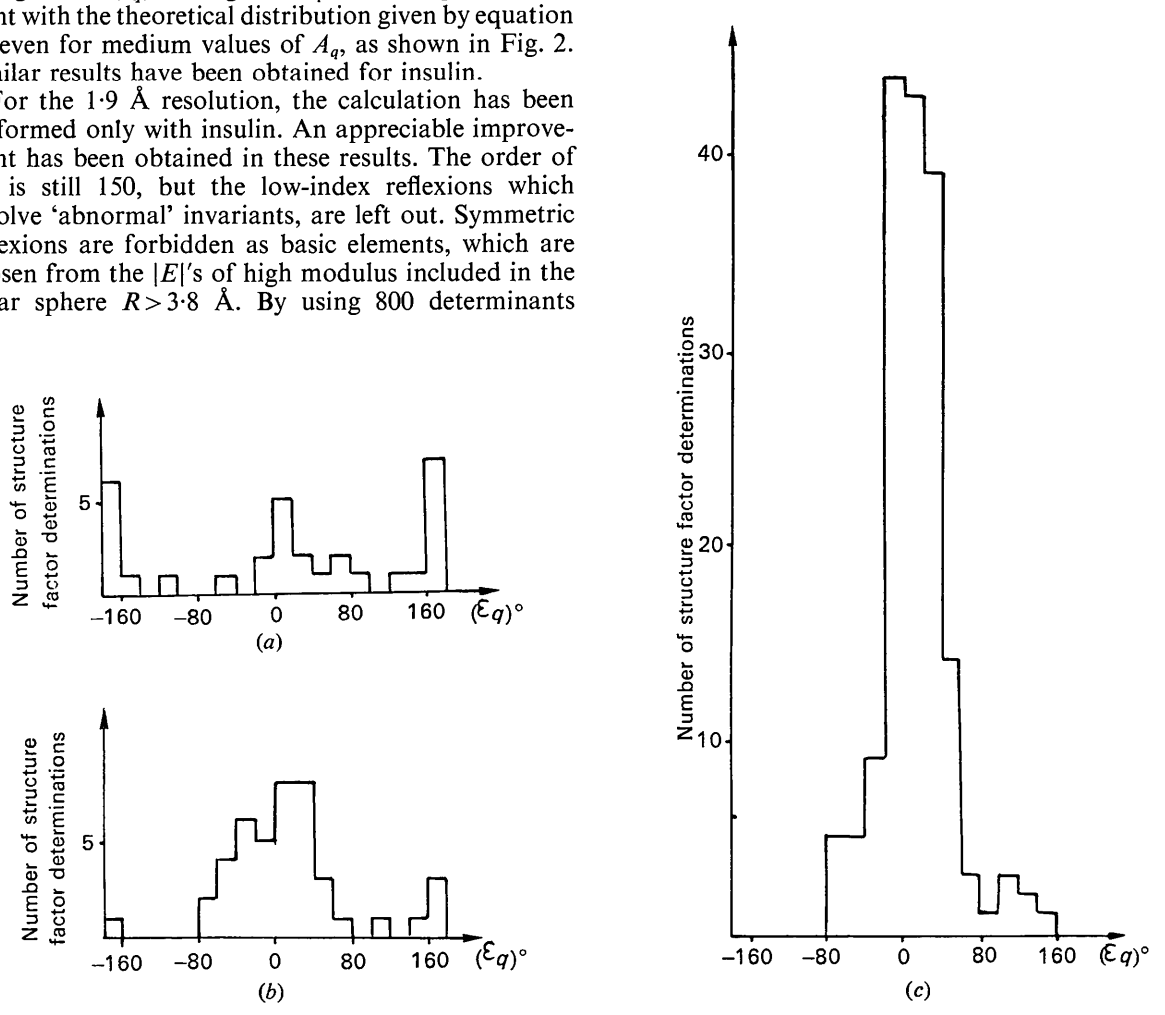

Fig. 6. Histograms of the errors in the phase determination of the reflexion 300 of insulin. (*a*) $m=100$; $800\Delta_{m+1}$ were considered; D_m construction criterion is based on $\sum|E_H^2|$; (*b*) $m=100$; 800 Δ_{m+1} were considered; D_m construction criterion is based on $\sum A\cos\varphi_{HK}$; (*c*) $m=400$; 500 Δ_{m+1} were considered; D_m construction criterion is based on $\sum A\cos\varphi_{HK}$.

Table 1. *Summary of results obtained from phase refinement for calculated data of insulin* (1·9 Å *resolution*)

		800 order-151 determinants		500 order-401 determinants	
	Total number of phases	Number of determined phases	$\Delta\Phi$	Number of determined phases	$\Delta\Phi$
$\|E\|>1$	2653	1264	21°	1800	15°
$\|E\|>1{\cdot}5$	836	525	15	653	12
$\|E\|>2$	214	147	13	178	10
$\|E\|>2{\cdot}5$	41	31	9	39	7

(Fig. 1); the distribution of $|\varepsilon_q|$, for a given A_q, is also in better agreement with the theoretical distribution (Fig. 2).

Influence of the order

An order-400 covariance matrix has been built up for insulin by choosing the basic elements according to the following criteria:

(*a*) The modulus must be higher than 1·30.

(*b*) Since the reflexions with low indices involve too many 'abnormal invariants', the basic-element reflexions are included in the region of the polar sphere $3{\cdot}8 < R < 6$ Å.

(*c*) Finally, we select in each step the basic element U_{1q} corresponding to the maximal value of the following expression S for all elements of the row:

$$S = \sum_{p=1}^{m} |U_{pq}U_{q1}U_{1p}| \cos(\varphi_{pq}+\varphi_{q1}+\varphi_{1p}) . \qquad (15)$$

The main characteristics of the determinants obtained by using these criteria are:

– The distribution of the structure-invariant phases involved in the D_m determinant is close to the theoretical distribution and, as a consequence, the variation of $\delta\varphi$ as a function of A_q remains in good agreement with the theoretical curve when the determinant order increases from 150 to 400 (Figs. 1 and 2).

– The value of the D_m determinant decreases as a function of the order more quickly than the theoretical mean value since the actual value of the ratio (Δ_{m+1}/D_m) is always lower than the mean value $(1-m/N)$, as shown in Fig. 3.

Using 500 order-401 determinants Δ_{m+1}, the phases of 1800 structure factors (with $|E|>1$) have been calculated with an average error $\Delta\Phi$ of 15°. The overall average value of $\Delta\Phi$ decreases when the determinant order increases as shown in Fig. 4. However, if too many errors are introduced in the phases and the moduli, the determinant may become negative and, of course, the above calculations lack any meaning.

The determination of the phases of the low-index reflexions is very sensitive to the characteristics and to the order of the D_m determinant. This fact is pointed out by the variation of the average value $\Delta\Phi$ which is given in Fig. 5, as a function of the order for reflexions corresponding to different resolutions. For all orders the higher value of $\Delta\Phi$ corresponds to the 6 Å resolution reflexions. The histograms of the differences ε_q corresponding to the reflexion 300 (Fig. 6) lead to similar remarks. Moreover, if the determinant is built up by using, as criterion, the maximal value of $(\sum|E_H^2|)$

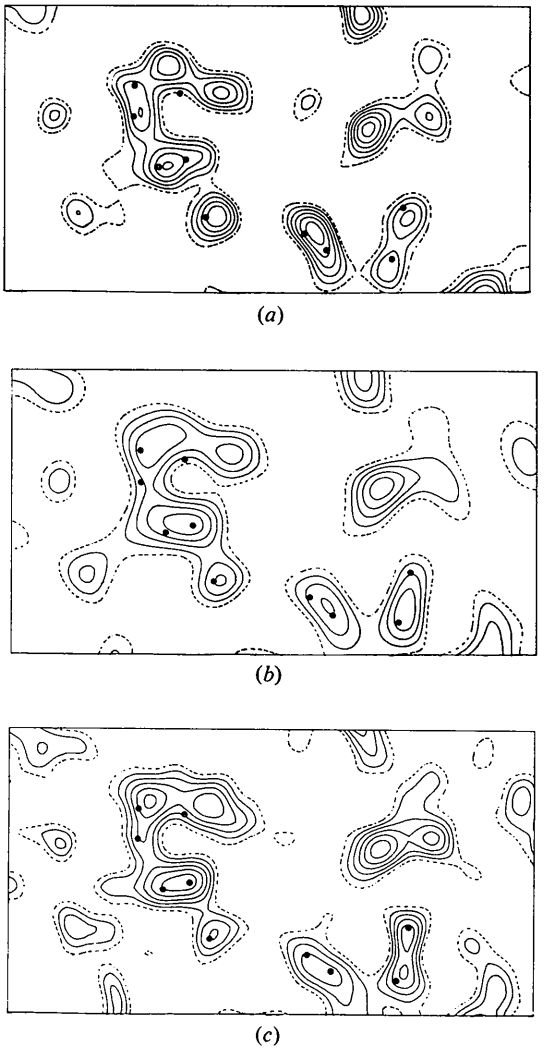

Fig. 7. A section of the insulin electron-density map calculated with: (*a*) phases evaluated by extension from 2·8 to 2 Å (two cycles of regression-equation calculations were performed); (*b*) isomorphous phases (2·8 Å resolution); (*c*) isomorphous phases (2 Å resolution). Atomic positions close to the section 2/48 *c* for residues B_{12}, B_{14} and B_{24} are drawn in according to the model structure.

Table 2. *Summary of results obtained from phase refinement and extension with a D_m determinant of order* 300, *for calculated data of insulin*

	Phase refinement (1·9 Å resolution)			Phase extension from 2·8 Å to 2·2 Å resolution		
	Total number of phases	Number of determined phases	$\Delta\Phi$	Total number of phases	Number of determined phases	$\Delta\Phi$
$\|E\|>1$	2653	1640	17°	893	621	23°
$\|E\|>1{\cdot}5$	836	627	14	215	203	19
$\|E\|>2$	214	172	13	41	39	19
$\|E\|>2{\cdot}5$	41	37	9	4	4	10

[instead of criterion (*c*) given above] for all elements of each row, some of the determined phases can be quite wrong, despite a large value of A_H, as shown Fig. 6(*a*).

Phase extension and refinement – application to actual structures

The above results have been used to initiate a method for phase refinement and for phase extension from medium to high resolution. The same criterion for the construction of D_m has been used as previously, but for the extension the moduli of structure factors having unknown phases have been set equal to zero (actually, a third of the structure factors involved). For the calculated data of insulin, the phases of 621 structure factors out of 893 structure factors have been evaluated by extension from 2·8 to 2·2 Å resolution, the phases for $R<2{\cdot}8$ Å being exact phases. These results, summarized in Table 2, have been obtained with an order-300 covariance matrix, after three cycles of regression-equation calculations. The average absolute value of the shift between the first and the second cycle is 31°; between the two last cycles it is 17°. This suggests a reasonable convergence for the method used. The phase-extension data are to be compared in Table 2 with those obtained by phase refinement.

The same method has been extended to the actual data of insulin. Here, we use the observed moduli; in the refinement process the starting phases, given by the isomorphous-replacement method, are included in the polar sphere $R>2$ Å. The phases of 1955 structure factors out of 2454 observable structure factors (with $|E|>1$) have been evaluated by using 500 order-401 determinants. The corresponding electron-density map has shown that the general configuration of the protein chain remains roughly unchanged; no interpretation of the significant changes has been attempted yet.

Similar calculations have been performed for the phase extension from 2·8 to 2 Å. An electron-density map [Fig. 7(*a*)] has been calculated with:

– for all reflexions included in the polar sphere $R>2{\cdot}8$ Å, the isomorphous phases;

– for the reflexions included in the region of the polar sphere $2{\cdot}8>R>2$ Å (with $|E|>1$), the phases evaluated by two cycles of regression-equation calculations.

This map seems in reasonable agreement with the electron-density map calculated with the isomorphous phases for 2·8 Å resolution [Fig. 7(*b*)]. On the other hand, some apparent differences occur between 'the phase-extension' map and the electron-density map calculated with isomorphous phases for 2 Å resolution [Fig. 7(*c*)]; nevertheless, the general configuration of the protein chain is still the same. The discussion of the stereochemically important details should be developed in the near future.

We are grateful to Professor D. C. Hodgkin and her colleagues for providing the data of insulin, and to Professor J. C. Kendrew and Professor D. Blow for allowing us to use the data of myoglobin. We thank Professor D. Sayre for his constant interest and fruitful discussions. We are indebted to the Centre Européen de Calculs Atomiques et Moléculaires for partial financial support.

References

BLUNDELL, T. L., CUTFIELD, J., DODSON, E. J., DODSON, G. G., HODGKIN, D. C., MERCOLA, D. & VIJAYAN, M. (1971). *Nature, Lond.* **231**, 506–511.

COCHRAN, W. (1955). *Acta Cryst.* **8**, 473–478.

KENDREW, J. C., DICKERSON, R. E., STRANBERG, B. E., HART, R. G., DAVIES, D. R., PHILLIPS, D. C. & SHORE, V. C. (1960). *Nature, Lond.* **185**, 422.

MAUGUEN, Y., RANGO, C. DE & TSOUCARIS, G. (1973). *Acta Cryst.* A**29**, 574–575.

RANGO, C. DE (1969). Thesis, Paris.

RANGO, C. DE & MAUGUEN, Y. (1972). CECAM Workshop Reports.

RANGO, C. DE, TSOUCARIS, G. & MAUGUEN, Y. (1973). Symposium on the Structures of Biological Molecules, Stockholm, Sweden.

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

SARRAZIN, M. (1970). CECAM Workshop Reports.

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