Lead Article

Acta Cryst. (1987). A43, 593-612

Direct Methods – from Birth to Maturity*

BY M. M. WOOLFSON

Department of Physics, University of York, York YO1 5DD, England

(*Received* 22 *October* 1986; *accepted* 8 *May* 1987)

Abstract

The award of the Nobel Prize for Chemistry to two pioneers of direct methods, H. Hauptman and J. Karle, has been a recognition of the importance that these methods have attained, in crystallography in particular and in science generally. The development of direct methods is traced from its first beginnings with Harker–Kasper inequalities and the Karle & Hauptman determinantal inequalities. The range of application of these methods to centrosymmetric structures was much increased by the introduction of the sign relationship by Sayre, Cochran and Zachariasen and an ACA Monograph by Hauptman & Karle to which the origins of the application of direct methods may be traced. Sayre's paper, in 1952, developed an exact equation which applied to both centrosymmetric and non-centrosymmetric structures as did the Karle & Hauptman determinantal inequalities. However, it was the derivation of the probability distribution for an individual three-phase relationship by Cochran in 1955 and the tangent formula by Karle & Hauptman in 1956 which provided the weaponry to tackle non-centrosymmetric structures, but nothing decisive was done until, in 1964, I. L. Karle & J. Karle solved the first non-centrosymmetric structure with direct methods using their symbolic addition procedure. The advent of the computer allowed automatic multisolution procedures, such as *MULTAN*, *SHELX* and *SIMPEL*, to take over as the main tools for the solution of small structures. Various developments since about 1970 have slowly introduced direct-methods concepts into the field of solving macromolecular structures, starting with the maximum-determinant method of Tsoucaris and progressing to the still experimental maximum-entropy method. Various predictions are made about the progress of direct methods in the next few years. A new tangent formula, the Sayre tangent formula, which gives most of the benefit of the maximum-entropy method but is much easier to apply, is suggested as a possible new source of progress. Other progress may be made in the association of direct methods with physical methods, such as isomorphous replacement and anomalous scattering, following work which has already been done over the past few years.

Prologue

The first experiments on the scattering of X-rays by a crystal, performed by Friedrich and Knipping in 1912 at the instigation of von Laue, were only concerned with exploring the periodicity in a crystal. Shortly afterwards, on the spectrometer built by W. H. Bragg, spectra were measured for some alkali halides and the Braggs solved these simple structures requiring only the concepts that various sets of atoms were scattering either in phase or π out of phase with each other for each of the diffracted beams.

In 1913 at a Solvay Conference devoted to the new subject of X-ray crystallography, Sommerfeld gave, in principle, the structure-factor equation

$$F(\mathbf{h}) = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j), \qquad (1)$$

where f_j is the scattering factor of the jth atom and $\mathbf{r}_j$ its vector position (with fractional coordinate components) in a unit cell containing N atoms.

The structure factor, $F(\mathbf{h})$, is a complex quantity,

$$F(\mathbf{h}) = |F(\mathbf{h})| \exp[i\varphi(\mathbf{h})], \qquad (2)$$

and the observed intensity $I(\mathbf{h})$ is proportional to the structure amplitude squared, $|F(\mathbf{h})|^2$, with the constant of proportionality depending on several physical factors – temperature, absorption, diffraction geometry *etc.* – whose effects can be estimated quite well.

From (1), if one is given the types and positions of atoms in the unit cell, it is possible to calculate the expected values of the observed quantities $I(\mathbf{h})$. However, what is required in practice is the solution of the inverse problem: if one is given the intensities, then how can the positions of the atoms be found? If the information available is a complete set of structure factors *in both magnitude and phase* then the problem is readily solved. In his Bakerian lecture to the Royal Society in 1915, W. H. Bragg suggested the use of the Fourier series in crystal-structure analysis. Thus the electron density at a point $\mathbf{r}$ in the unit cell

* *Editorial note*: This invited paper is one of a series of comprehensive Lead Articles which the Editors invite from time to time on subjects considered to be timely for such treatment.

0108-7673/87/050593-20$01.50

is given by

$$\rho(\mathbf{r}) = (1/V) \sum_{\mathbf{h}} F(\mathbf{h}) \exp(-2\pi i \mathbf{h} \cdot \mathbf{r}) \qquad (3)$$

or, in an alternative form,

$$\rho(\mathbf{r}) = (1/V) \sum_{\mathbf{h}} |F(\mathbf{h})| \cos[2\pi \mathbf{h} \cdot \mathbf{r} - \varphi(\mathbf{h})], \qquad (4)$$

where the summation is over all observed reflexions.

Experiments can give the structure amplitudes, $|F(\mathbf{h})|$, but not the phases, $\varphi(\mathbf{h})$, and this constitutes the phase problem in crystallography. The earliest attempts to solve the phase problem were by trial-and-error methods - essentially guessing a structure and then seeing whether calculated intensities agreed with those observed. At best this could only be applied to simple structures where some outside information was available - for example, from symmetry.

The Patterson function (Patterson, 1934) not only provided an important new tool for structure analysis but indicated that, in principle at any rate, the data alone were capable of giving the structure. Multiplication of each side of (1) by its own complex conjugate gives

$$|F(\mathbf{h})|^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i f_j \exp[2\pi i(\mathbf{r}_i - \mathbf{r}_j)], \qquad (5)$$

and by the same process that relates (4) to (1) we can derive a Fourier (Patterson) map

$$P(\mathbf{r}) = (1/V) \sum_{\mathbf{h}} |F(\mathbf{h})|^2 \cos(2\pi \mathbf{h} \cdot \mathbf{r}), \qquad (6)$$

which gives a peak at each interatomic position $\mathbf{r}_i - \mathbf{r}_j$ with a weight proportional to $Z_i Z_j$. It was shown by Wrinch (1939) that knowledge of the complete vector set would automatically reveal the structure. In practice a Patterson map is usually so overcrowded that individual peaks cannot all be recognized but with an infinite amount of data, *i.e.* infinite resolution, they could be found and hence yield the structure.

In cases where the structure contains a few heavy atoms, or high symmetry, then associated peaks can be found and, with the heavy-atom or symmetrical-group coordinates available, the whole structure can usually be solved. Another procedure, which also uses the Patterson map, is that of isomorphous replacement where two structures differ only in that one atom, or group of atoms, in one of them is replaced by a different kind of atom, or group, in the other. The differences in the magnitudes of structure factors, or intensities, from the two structures can be used as coefficients in a Patterson map to show the position of the isomorphously-replaced atoms. Isomorphism can be obtained with protein structures by chemically attaching residues carrying heavy atoms (*e.g.* Hg) at various points within them. The multiple-isomorphous-replacement method, which uses several sets of isomorphous data, is largely responsible for the many triumphs of protein structural crystallography.

There were early attempts to obain interatomic vectors directly from (5), in 1927 by Ott and in 1938 by Avrami, but their methods could only be applied to trivial problems. However, the term 'direct methods' is usually taken to mean that class of methods which attempt by mathematical means to derive the *phases* of the structure factors using only the intensity information. Here we shall be describing the development of such methods from their earliest beginnings to their present comparatively advanced state. It should be said in advance that this survey will not be an exhaustive one; developments which seemed significant at the time, especially some of the earlier ones, have proved to be of little value and to describe them in detail would tend to be distracting rather than instructive. Instead, the major thread of the development of direct methods will be traced over the years and the history of the gradual evolution of these methods from being a primitive aid for the solution of simple structures to their present dominant role in structural crystallography will be related.

In 1985 two pioneers of direct methods, J. Karle and H. Hauptman, were suitably honoured by the award of the Nobel Prize for Chemistry. In particular, the crucial role of their work in the successful growth of this field of science will be highlighted in the account which follows.

The first few steps: 1948–1951

It was perhaps appropriate that the published paper which heralded the birth of direct methods appeared in the first issue of *Acta Crystallographica*, the journal of the newly-formed International Union of Crystallography which now carries so much of the fruit of that seedling. This paper, by Harker & Kasper (1948), presented inequality relationships between structure factors which, in some cases, could give unambiguous phase information.

For the purposes of direct methods the normal structure factor is not the best one to use. It is better to consider structure factors corresponding to point atoms (electron density a δ function) with no thermal motion. The scattering factor for such an atom is the same for every reflexion and there is no tendency for the resultant structure factors systematically to fall off with increasing scattering angle. Two such structure factors are in common use. The first is the *unitary structure factor*, $U(\mathbf{h})$, which satisfies the relationships

$$\langle |U|^2 \rangle = \sum_{j=1}^{N} n_j^2 \qquad (7a)$$

where

$$n_j = f_j \Big/ \sum_{j=1}^{N} f_j$$

is the *unitary scattering factor* and

$$0 \le |U| \le 1. \qquad (7b)$$

The second kind, and latterly the more important, is the normalized structure factor, $E(\mathbf{h})$, which satisfies the relationships

$$\langle|E|^2\rangle = 1 \tag{8a}$$

and, for equal atoms,

$$E(\mathbf{h}) = N^{1/2}U(\mathbf{h}). \tag{8b}$$

Harker & Kasper used well-known mathematical inequalities, mainly the Cauchy inequality,

$$\left|\sum_{j=1}^{N} a_j b_j\right|^2 \leq \sum_{j=1}^{N} |a_j|^2 \times \sum_{j=1}^{N} |b_j|^2, \tag{9}$$

together with the structure-factor equations which, with space-group information included, could take on forms other than that shown in (1). Thus, for space group $P\bar{1}$ containing just a centre of symmetry,

$$U(\mathbf{h}) = \sum_{j=1}^{N} n_j \cos(2\pi\mathbf{h}\,.\,\mathbf{r}_j), \tag{10}$$

and by applying the Cauchy inequality in different ways one finds

$$U(\mathbf{h})^2 \leq \tfrac{1}{2}[1 + U(2\mathbf{h})] \tag{11a}$$

and

$$[U(\mathbf{h}) \pm U(\mathbf{k})]^2 \leq [1 \pm U(\mathbf{h}+\mathbf{k})][1 \pm U(\mathbf{h}-\mathbf{k})]. \tag{11b}$$

For a centrosymmetric structure with the centre of symmetry at the origin the structure factors are real and, instead of the *phase* $\varphi(\mathbf{h})$, we can think of the *sign*, $s(\mathbf{h})$, of the structure factor, where $\varphi(\mathbf{h}) = 0$ corresponds to $s(\mathbf{h}) = +1$ and $\varphi(\mathbf{h}) = \pi$ corresponds to $s(\mathbf{h}) = -1$. In terms of signs, where all the involved unitary structure factors are non-zero, inequality 11(b) can be written

$$\begin{aligned}[|U(\mathbf{h})| + |U(\mathbf{k})|]^2 &\leq [1 + s(\mathbf{h})s(\mathbf{k})s(\mathbf{h}+\mathbf{k})|U(\mathbf{h}+\mathbf{k})|] \\ &\quad \times [1 + s(\mathbf{h})s(\mathbf{k})s(\mathbf{h}-\mathbf{k})|U(\mathbf{h}-\mathbf{k})|],\end{aligned} \tag{12}$$

and if the $|U|$'s are sufficiently large then it may be shown that

$$s(\mathbf{h})s(\mathbf{k})s(\mathbf{h}+\mathbf{k}) = 1 \tag{13a}$$

and/or

$$s(\mathbf{h})s(\mathbf{k})s(\mathbf{h}-\mathbf{k}) = 1. \tag{13b}$$

Relationships (13) did not appear in the original work of Harker & Kasper (1948), although they were implicit in their results.

Similarly, if $|U(\mathbf{h})|$ and $|U(2\mathbf{h})|$ are large enough then it can be shown from (11a) that $s(2\mathbf{h})$ is positive.

The drawback of these inequality relationships is that they are only useful for structures with few atoms. For N equal atoms in the unit cell then $\langle|U|^2\rangle = 1/N$ and only few reflexions will have $|U|$ with value greater than three times the r.m.s. value of U. Thus with $N = 64$ there will be a few $|U|$ values greater than 0·4 and not many, if any, inequality relationships will be found.

Other types of algebraic inequality formulae have been used to give inequality relationships between structure factors, notably by Gillis (1948), Okaya & Nitta (1952) and Sakurai (1952). However, they are never more powerful than the basic Harker-Kasper inequalities and have been little used.

Two years after the Harker-Kasper paper there appeared an important paper by Karle & Hauptman (1950) on determinantal inequalities. The underlying basis for this was quite old (Toeplitz, 1911) but, in a familiar form, it gives the result that for a Fourier summation to give a function everywhere non-negative then, for any order of determinant, $n+1$,

$$\begin{vmatrix} F(\mathbf{0}) & F(-\mathbf{h}_1) & F(-\mathbf{h}_2) & \cdots & F(-\mathbf{h}_n) \\ F(\mathbf{h}_1) & F(\mathbf{0}) & F(\mathbf{h}_1-\mathbf{h}_2) & \cdots & F(\mathbf{h}_1-\mathbf{h}_n) \\ F(\mathbf{h}_2) & F(\mathbf{h}_2-\mathbf{h}_1) & F(\mathbf{0}) & & F(\mathbf{h}_2-\mathbf{h}_n) \\ \vdots & \vdots & \vdots & & \vdots \\ F(\mathbf{h}_n) & F(\mathbf{h}_n-\mathbf{h}_1) & F(\mathbf{h}_n-\mathbf{h}_2) & \cdots & F(\mathbf{0}) \end{vmatrix} \geq 0, \tag{14}$$

where the terms in the leading column must be different (but can be symmetry-related) structure factors. It will be noted that the elements form an Hermitian matrix and hence the value of the determinant is real. Since the U's (and E's) correspond to point atoms then the inequality can be written in terms of these quantities.

One order-three determinant which can be formed, with $U(\mathbf{0}) = 1$, is

$$\begin{vmatrix} 1 & U(-\mathbf{h}) & U(-2\mathbf{h}) \\ U(\mathbf{h}) & 1 & U(-\mathbf{h}) \\ U(2\mathbf{h}) & U(\mathbf{h}) & 1 \end{vmatrix} \geq 0, \tag{15a}$$

which, when expanded, gives the Harker-Kasper inequality (11a). Another determinant is

$$\begin{vmatrix} 1 & U(-\mathbf{h}_1) & U(-\mathbf{h}_2) \\ U(\mathbf{h}_1) & 1 & U(\mathbf{h}_1-\mathbf{h}_2) \\ U(\mathbf{h}_2) & U(\mathbf{h}_2-\mathbf{h}_1) & 1 \end{vmatrix} \geq 0, \tag{15b}$$

which gives for a centrosymmetric structure

$$1 - U(\mathbf{h}_1)^2 - U(\mathbf{h}_2)^2 - U(\mathbf{h}_1-\mathbf{h}_2)^2 + 2U(\mathbf{h}_1)U(\mathbf{h}_2)U(\mathbf{h}_1-\mathbf{h}_2) \geq 0. \tag{15c}$$

This was the first explicit appearance of the product of three signs which played so dominant a role in subsequent development.

For large U magnitudes this can yield the conclusion (13b) but, it will be noticed, not in conjunction with (13a), as it can be with the Harker-Kasper inequality.

To obtain a greater insight into the information contained in the inequality (15b) it can be recast in

a form given by Karle & Hauptman (1950):

$$|U(\mathbf{h}_1)-U(\mathbf{h}_2)U(\mathbf{h}_1-\mathbf{h}_2)| \leq \begin{vmatrix} 1 & U(-\mathbf{h}_2) \\ U(\mathbf{h}_2) & 1 \end{vmatrix}^{1/2} \begin{vmatrix} 1 & U(\mathbf{h}_2-\mathbf{h}_1) \\ U(\mathbf{h}_1-\mathbf{h}_2) & 1 \end{vmatrix}^{1/2} \quad (16a)$$

which is of the form

$$|U(\mathbf{h}_1)-\delta| \leq r. \quad (16b)$$

The meaning of (16*b*) can be seen in Fig. 1, which shows the possible values of $U(\mathbf{h}_1)$ in the complex plane. This puts a restriction on the phase of $U(\mathbf{h}_1)$ and, if r is small, we find that

$$\varphi(\mathbf{h}_1) \simeq \varphi(\mathbf{h}_2)+\varphi(\mathbf{h}_1-\mathbf{h}_2). \quad (17)$$

For a reasonably complicated structure one would normally find the point C close to the origin and the radius of r so large that no phase angle, $\varphi(\mathbf{h}_1)$, is excluded. However, as we shall see later even this does not completely negate the relationship (17).

In their 1950 paper Karle & Hauptman showed that the form of expression (16*b*) can also be found for high-order determinants. The values of δ and r then become functions of determinants found from the main determinant by moving rows and columns - and other changes. All that need be noted here is that if a large body of phase information is already available so that δ and r can be evaluated then, in principle, r can be very small, or even zero, and $\varphi(\mathbf{h}_1)$ may then be determined with precision.

There is a great deal of information inherent in the determinantal inequalities and much subsequent development in the theory of direct methods has been found to be an aspect of what they contain - albeit that the connection is not always blindingly obvious.

We have already seen that an inequality relationship is capable of restricting the range of values of a general phase without restricting it to a particular value, as in (17). For a centrosymmetric structure the only possible phases can be represented by $s(\mathbf{h}) = +1$ or -1 so that the idea of restricting to a range does not apply. However, Gillis (1948) made the point that, even when the structure factors were not of sufficiently large magnitude to define the sign of a particular structure factor or product of structure factors, there might still be an implication, from the near-validity of the inequality, that the sign has a particular value. We shall see that this is true and we shall now follow the development of probability relationships between structure factors.

Setting the stage: 1952–1956

Some very simple centrosymmetric structures could be solved by Harker-Kasper inequalities but by and large they made little impact in the practical area of solving crystal structures. The Patterson function still reigned supreme, and, since most calculations were being done by hand, using Beevers-Lipson strips at best, nearly all structures were being solved from two-dimensional projections. It should be mentioned here that a critical element of thinking, which added greatly to the impetus in developing direct methods, was the realization that the problem was greatly overdetermined. To the condition of non-negativity there could be added that of atomicity; with this the number of observations exceeds the number of degrees of freedom and in principle, the problem becomes soluble (Hauptman & Karle 1950*a*, *b*). Three important papers which appeared in the same issue of *Acta Crystallographica* - by Sayre (1952), Cochran (1952) and Zachariasen (1952) - incorporated the concept of atomicity and heralded a new stage in the development of direct methods. The paper by Sayre was the most fundamental. Sayre considered the result of squaring the electron density of a structure consisting of equal resolved atoms. This is shown in Fig. 2; it is clear that the squared structure resembles the original structure in having equal peaks in the same

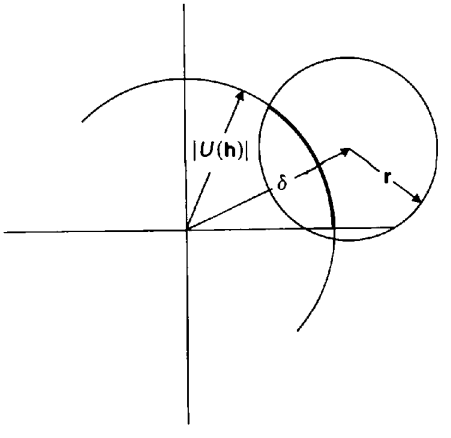

Fig. 1. The Karle-Hauptman determinant inequality indicates that $U(\mathbf{h})$ must be somewhere on the thickened arc.

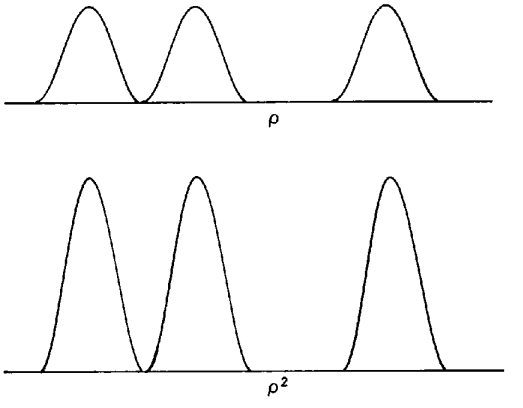

Fig. 2. The basis of the Sayre equation. For equal atoms ρ and ρ^2 both show equal resolved regions of density.

positions but with peaks of different shape. Using the well known theorem that the Fourier transform of a product of two functions is the convolution of the Fourier transforms of the individual functions Sayre arrived at the expression

$$F(\mathbf{h}) = \theta(\mathbf{h}) \sum_{\mathbf{k}} F(\mathbf{k}) F(\mathbf{h}-\mathbf{k}), \tag{18}$$

where $\theta(\mathbf{h})$ can be determined for atoms of known shape. Equation (18) is known as Sayre's equation and relates the structure factors exactly if, and only if, the structure consists of equal resolved atoms.

At first glance it seems that Sayre's equation would lead directly to the phases of structure factors if in some way a set of phases could be found which makes all the equations hold. In his paper Sayre actually found the signs of structure factors for a centrosymmetric projection of hydroxyproline, but it was not a straightforward task and was not helped by the fact that the projection did not consist of equal resolved atoms.

The approach by Cochran (1952) was a little more intuitive. He argued that a correct electron-density distribution would contain large near-zero regions with the density concentrated around the atomic positions. He interpreted this as having the quantity

$$\int_V \rho^3 \, \mathrm{d}V \quad \text{large and positive,} \tag{19}$$

where the integration is taken over the whole unit cell. It may be shown that (19) is equivalent to having a large value of

$$Q = \sum_{\mathbf{h}} \sum_{\mathbf{k}} F(\mathbf{h}) F(\mathbf{k}) F(-\mathbf{h}-\mathbf{k}). \tag{20}$$

For a centrosymmetric structure the value of Q will be large if, in general, the contributors to the summation are positive, and this led Cochran to the triple-product sign relationship (TPSR)

$$s(\mathbf{h}) s(\mathbf{k}) s(\mathbf{h}+\mathbf{k}) \approx +1, \tag{21}$$

where $\approx$ means 'probably equals'.

Zachariasen (1952) also arrived at relationship (21), although by somewhat flawed reasoning. An important contribution made by him was to demonstrate an effective method of using the TPSR to solve the structure of metaboric acid.

Zachariasen used letter symbols to represent unknown signs and by applying inequality relationships to three-dimensional data he was able to represent 40 signs in terms of five letter symbols. He then applied an extension of relationship (21),

$$s(\mathbf{h}) \approx s\left[\sum_{\mathbf{k}} s(\mathbf{k}) s(\mathbf{h}-\mathbf{k})\right], \tag{22}$$

where the terms on the right-hand side of (22) were pairs of known signs giving an indication for $s(\mathbf{h})$. Thus if the contributors on the right-hand side were

$$ab \; ab \; ab \; ab \; ab \; cd \; cd \; cd +$$

then it would be deduced that $s(\mathbf{h}) \approx ab$, that $abcd$ is very probably positive and that the products ab and cd might be positive. By this kind of process Zachariasen was able to extend the knowledge of signs and to find a unique set for the 168 largest $|U|$ values. By an exactly similar process Cochran & Penfold (1952) solved the structure of L-glutamine.

In the following year Hauptman & Karle (1953) produced a monograph which was to be quite influential - even if its title was somewhat optimistic! In this monograph they introduced the normalized structure factor E, to which reference has already been made. The distribution of $|E|$'s is independent of structural complexity and the theory and practice of direct methods is greatly simplified by their use. The main conclusions they reached was that the sign of $E(\mathbf{h})$ was given by $\sum_1 + \sum_2 + \sum_3 + \sum_4$, where

$$\sum_1 = (\sigma_3 / 4\sigma_2^{3/2}) \sum_{\mathbf{h}_\mu = \mathbf{h}/2} [E(\mathbf{h}_\mu)^2 - 1] \tag{23a}$$

$$\sum_2 = (\sigma_3 / 2\sigma_2^{3/2}) \sum_{\mathbf{h}_\nu + \mathbf{h}_\mu = \mathbf{h}} E(\mathbf{h}_\nu) E(\mathbf{h}_\mu) \tag{23b}$$

$$\sum_3 = (\sigma_4 / 4\sigma_2^2) \sum_{\mathbf{h}_\mu + 2\mathbf{h}_\nu = \mathbf{h}} E(\mathbf{h}_\mu)[E(\mathbf{h}_\nu)^2 - 1] \tag{23c}$$

$$\sum_4 = (\sigma_5 / 8\sigma_2^{5/2}) \sum_{2\mathbf{h}_\mu + 2\mathbf{h}_\nu = \mathbf{h}} [E(\mathbf{h}_\mu)^2 - 1][E(\mathbf{h}_\nu)^2 - 1] \tag{23d}$$

and $\sigma_n = \sum_{j=1}^{N} z_j^n$, where z_j is the atomic number of the jth atom.

Although these equations were derived by probability theory they do have interpretations in terms of physical functions. The quantity $|E(\mathbf{h})|^2 - 1$ is the Fourier coefficient of a Patterson function with its origin peak removed and the four summations, if all terms are included on the right-hand side, are related respectively to the Patterson function, the squared electron density, the product of electron density with a half-scale Patterson function and the square of the Patterson function.

Associated with these formulae, Hauptman & Karle derived probability expressions but the final approximation they presented did not constrain the probability that $s(\mathbf{h})$ was positive within the necessary limits of zero and unity. The first probability formula for a single TPSR, which was valid within the limitations imposed by use of the central-limit theorem and only for equal-atom structures, was given by Woolfson (1954), but a formula valid for non-equal atoms was given by Cochran & Woolfson (1955) in the form

$$P_+(\mathbf{h}, \mathbf{k}) = \tfrac{1}{2} + \tfrac{1}{2} \tanh [(\varepsilon_3 / \varepsilon_2^3) |U(\mathbf{h}) U(\mathbf{k}) U(\mathbf{h}+\mathbf{k})|] \tag{24}$$

where

$$\varepsilon_n^* = \sum_{j=1}^{N} n_j^n.$$

The same workers found an expression that $s(\mathbf{h})$ is positive, given many pairs of contributors on the right-hand side of (23*b*), in the form

$$P_+(\mathbf{h}) = \tfrac{1}{2} + \tfrac{1}{2}\tanh\left[(\varepsilon_3/\varepsilon_2^3)|U(\mathbf{h})|\sum_{\mathbf{k}} U(\mathbf{k})U(\mathbf{h}-\mathbf{k})\right]. \tag{25}$$

These equations are valid only for structure factors with magnitudes well away from the region where inequalities would hold; a much more precise analysis, giving formulae valid over a much larger range, was given by Klug (1958).

During the years this theory was being developed the first practical computers were becoming available. Cochran & Douglas (1955) designed the first direct method to run on a computer. The machine they had available was the EDSAC I which offered a store of 1024 17-bit words, no backing store, no floating-point capacity, an add time of 1 ms and a multiplication time of 3 ms. Because of its historical interest we shall briefly look at their algorithm. They started with 20 structure factors for a projection (*pgg*) of salicylic acid; two signs could be chosen to fix the origin and two others were fixed by the equivalent of Hauptman & Karle's $\sum_1$ formula. The 20 structure factors, the unknown signs of which were indicated by x_1 to x_{16}, were linked by 29 TPSR's whose signs were indicated by s_1 to s_{29}.

Equations could be found of the form

$$s_1 = -x_1x_9x_{15}, \quad s_2 = x_1x_2x_{12} \quad \textit{etc.},$$

and by a suitable selection of the TPSR's it was possible to express each x as a product involving only s_1 to s_{16} (set *A*), thus:

$$\begin{aligned} x_1 &= s_1s_3s_4s_5s_7s_8s_9s_{10}s_{12}s_{13}s_{15}s_{16} \\ x_2 &= -s_1s_2s_3s_4s_8s_9s_{11}s_{12}s_{15}s_{16} \\ &\vdots \\ x_{16} &= s_1s_2s_3s_5s_7s_9s_{10}s_{13}s_{15}s_{16}, \end{aligned} \tag{26a}$$

and the remaining s values (set *B*) also in terms of s_1 to s_{16} appear as

$$\begin{aligned} s_{17} &= s_4s_6 \\ s_{18} &= s_{11}s_{13}s_{14} \\ &\vdots \\ s_{29} &= s_1s_2s_3s_5s_7s_9s_{10}s_{13}s_{15}s_{16}. \end{aligned} \tag{26b}$$

On the basis of probability formula (24) it was concluded that not more than three of set *A* should be negative and not more than five of the total set *A* + set *B*. First it is assumed that s_1 to s_{16} are all positive and these signs are substituted in (26*b*). If five or less of these 13 s's are negative then the solution is acceptable and the x's found from (26*a*). Next are considered, one at a time, the 16 ways in which one of the members of set *A* can be negative, and if one of these leads to four or less of set *B* being negative then again an acceptable set of signs is found. Exploring up to three failures in set *A* required the examination of 697 possibilities and 24 sets of signs were acceptable under the criteria being used.

The Cochran & Douglas (1955) procedure was the first multiple-solution method which systematically sought for a number of plausible sets of signs. While it is possible in principle to choose between them by computing electron-density maps and recognizing the correct structure it is better, and more economical, to find some figure of merit (FOM) to rank the solutions in order of plausibility.

Cochran & Douglas examined two FOM's. The first of these was the value of

$$X = \sum_{\mathbf{h}}\sum_{\mathbf{k}} U(\mathbf{h})U(\mathbf{k})U(\mathbf{h}+\mathbf{k}) \tag{27}$$

where the summation was over all TPSR's. It is a measure of how well the TPSR's hold but, in the very nature of the sign-determining process, X will be large for all plausible sets of signs; for salicylic acid X for the correct set was ninth largest out of the 24 values.

A better figure of merit was found to be the 'zero check'. This was the value of

$$Z_0 = \sum_{\mathbf{h}}\left|\sum_{\mathbf{k}} U(\mathbf{k})U(\mathbf{h}-\mathbf{k})\right|, \tag{28}$$

where the terms of the summation were derived from the 20 structure factors whose signs were to be determined and the outer summation was over a number of values of $\mathbf{h}$ for which $|U(\mathbf{h})|$ was zero or very small. The Z_0 figure of merit is related to Sayre's equation, for if all terms were included then the $\mathbf{k}$ summation would be zero, or small, for a correct set of signs. A low value of Z_0 is therefore expected, and it was found that it was a very effective FOM.

In the early years of the application of direct methods attention was very much fixed on the centrosymmetric problem, and, although relationships between general phases were implicit in many of the inequalities and other formulae being produced [*e.g.* equation (17), derived from the Karle & Hauptman determinantal inequalities], there was no clear statement of how general phases might be related. Eventually this was given by Cochran (1955). The form was that the quantity

$$\Phi_3(\mathbf{h},\mathbf{k}) = \varphi(\mathbf{h}) + \varphi(\mathbf{k}) + \varphi(\bar{\mathbf{h}}+\bar{\mathbf{k}}) \tag{29a}$$

has a probability distribution

$$P[\Phi_3(\mathbf{h},\mathbf{k})] = \frac{\exp\{-\kappa(\mathbf{h},\mathbf{k})\cos[\Phi_3(\mathbf{h},\mathbf{k})]\}}{2\pi I_0[\kappa(\mathbf{h},\mathbf{k})]} \tag{29b}$$

where

$$\kappa = 2(\sigma_3/\sigma_2^{3/2})|E(\mathbf{h})E(\mathbf{k})E(\bar{\mathbf{h}}+\bar{\mathbf{k}})| \tag{29c}$$

and I_0 is a modified Bessel function of the second kind.

An alternative form of (29a), obtained by changing the form of indices somewhat, is

$$\varphi(\mathbf{h}) \approx \varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k}), \qquad (30)$$

and then the probability distribution, shown in Fig. 3, is for the values of $\varphi(\mathbf{h})$. The general form of the distribution is that as κ increases so the variance of the distribution decreases, *i.e.* the value becomes more highly constrained near the expectation value.

The concluding step in this setting-of-the-stage era was again provided by Karle & Hauptman (1956). Relationship (30) gives a probable value for $\varphi(\mathbf{h})$ when there is a pair of known phases $\varphi(\mathbf{k})$ and $\varphi(\mathbf{h}-\mathbf{k})$. What was now required was an estimate for $\varphi(\mathbf{h})$ when several pairs of known phases were available. This was given by Karle & Hauptman (1956) in the form of the tangent formula which, in a slightly modified form, is

$$\tan[\varphi(\mathbf{h})] \approx \frac{\sum_{\mathbf{k}} \kappa(\mathbf{h},\mathbf{k}) \sin[\varphi(\mathbf{k})+\varphi(\mathbf{h}-\mathbf{k})]}{\sum_{\mathbf{k}} \kappa(\mathbf{h},\mathbf{k}) \cos[\varphi(\mathbf{k})+\varphi(\mathbf{h}-\mathbf{k})]}. \qquad (31)$$

It is interesting to note that Sayre (1952), in developing signs by means of (18), carried out processes of looking at large TPSR's and gradually extending knowledge of signs which is paralleled in methods used at the present time.

Crystallographers had now been given the tools but they were a long way from finishing the job.

The doldrum years: 1957–1962

After five years of spectacular progress there began a period, not completely barren by any means but comparatively so in contrast to what had preceded it. Such activity as there was mostly consisted of devising methods of using the TPSR to solve centrosymmetric structures or projections. For example

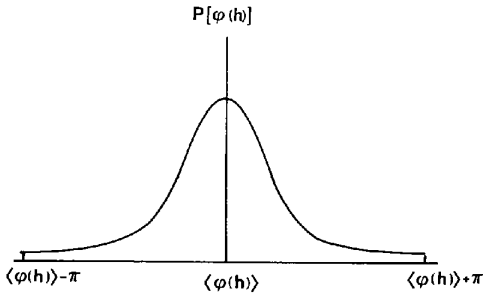

Fig. 3. The distribution function for $\varphi(\mathbf{h})$ given $\varphi(\mathbf{k})$, $\varphi(\mathbf{h}-\mathbf{k})$ and the magnitudes of the three normalized structure factors. $\langle\varphi(\mathbf{h})\rangle = \varphi(\mathbf{k})+\varphi(\mathbf{h}-\mathbf{k})$.

for the two-dimensional space group *pgg* there could be pairs of sign relationships of the form

$$s(4\,3\,0)s(3\,\bar{1}\,0) \approx s(7\,2\,0)$$

and

$$s(4\,3\,0)s(3\,\bar{1}\,0) \approx s(1\,4\,0)$$

or

$$s(4\,\bar{1}\,0)s(3\,3\,0) \approx s(7\,2\,0)$$

and

$$s(4\,\bar{1}\,0)s(3\,3\,0) \approx -s(1\,4\,0).$$

The conclusions drawn from each pair of relationships are contradictory but if, say, the magnitudes of the structure factors favour the first pair rather than the last then it may be deduced that

$$s(7\,2\,0) = s(1\,4\,0).$$

A much stronger situation would arise with one index in common for this space group.

Consider the pair of relationships

$$s(h\,k'\,0)s(h'\,k+k'\,0) \approx s(h'-h\,k\,0)$$

and

$$s(h\,k'\,0)s(h'\,k+k'\,0) \approx (-1)^{h+k'}s(h'+h\,k\,0).$$

This leads to the result

$$s(h'-h\,k\,0) \approx (-1)^{h+k'}s(h'+h\,k\,0),$$

and several different k' can be used. If one parity of k' gives much stronger indications than the other then a clear relationship between $s(h'-h\,k\,0)$ and $s(h'+h\,k\,0)$ may be evident (Grant, Howells & Rogers, 1957). A somewhat related method was devised by Woolfson (1958), but neither method could be regarded as decisive in advancing the effectiveness of direct methods. During this period small structures were being solved by the Zachariasen procedure and Woolfson (1961) proposed what he called the 'hit and miss' method. This was based on the observation that the solution of L-glutamine by Cochran & Penfold (1952) had begun with the determination of signs, in terms of symbols, for 13 of the largest structure factors from inequality relationships. However, when the structure was solved and structure factors were calculated an interesting fact emerged. They had overestimated their U's, especially those at high angle, and not one of their assumed inequality relationships was valid. Woolfson therefore suggested that a way of proceeding would be to assume that a few of the strongest sign relationships were inviolable and symbols could then be used in the usual way to extend sign information.

A small development, which does influence modern direct methods to some extent, was the introduction of the E map by Karle, Hauptman, Karle & Wing

(1958). This is a Fourier map with E replacing F and it has the effect of giving somewhat sharper peaks and enhancing the contribution of high-order reflexions whose phases have been estimated. There also results a considerable amount of background ripple, but this is in the region between the main peaks and causes no problems. An intrinsic part of most modern direct methods involves computation of E maps.

A great deal of the early work of Karle & Hauptman was concerned with the estimation of structure invariants and structure semi-invariants (Hauptman & Karle, 1956, 1959; Karle & Hauptman, 1961). The former are quantities which are independent of the choice of origin in the unit cell, examples of which are $|E(\mathbf{h})|$, $E(\mathbf{h})E(\bar{\mathbf{k}})E(\bar{\mathbf{h}}+\mathbf{k})$ or the associated quantity $\varphi(\mathbf{h})-\varphi(\mathbf{k})-\varphi(\mathbf{h}-\mathbf{k})$. Structure semi-invariants are quantities which do not change value by transfer from one special origin to another - *i.e.* from one centre of symmetry to another in $P\bar{1}$ or one of the eight origins in $P2_12_12_1$ equivalently related to the symmetry elements. Examples of these are, for $P\bar{1}$, $E(2\mathbf{h})$, $E(\mathbf{h}+\mathbf{k})E(\mathbf{h}-\mathbf{k})$ and, for $P2_12_12_1$, $\varphi(\mathbf{h}_1)+\varphi(\mathbf{h}_2)+\varphi(\mathbf{h}_3)$, where $\mathbf{h}_1+\mathbf{h}_2+\mathbf{h}_3=(0\,0\,0)$ modulo(2 2 2). This work is crucial to origin and enantiomorph specification on which many methods depend for initiating phase determination.

A formula for estimating the sum of three phases structure-invariant for $P1$ was given by Karle & Hauptman (1957) in the form, for the equal-atom case,

$$\begin{aligned}|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|\cos[\varphi(\mathbf{h})-\varphi(\mathbf{k})-\varphi(\mathbf{h}-\mathbf{k})]\\ \approx(N^{3/2}/2)\langle[|E(\mathbf{H})|^2-1][|E(\mathbf{h}+\mathbf{H})|^2-1]\\ \times[|E(\mathbf{h}-\mathbf{k}+\mathbf{H})|^2-1]\rangle_{\mathbf{H}}\\ +N^{-1/2}[|E(\mathbf{h})|^2+|E(\mathbf{k})|^2+|E(\mathbf{h}-\mathbf{k})|^2-2].\end{aligned} \tag{32}$$

In this equation the average is for $\mathbf{H}$ ranging over the whole of reciprocal space, and it would seem to be a way of involving the whole data set in the determination of a single triple-phase invariant. This equation turns out to be somewhat unreliable and, indeed, the magnitude of the right-hand side can easily exceed the magnitude of the left-hand side by a large factor. The equation is clearly related to the Patterson function and would be valid if the Patterson map had all peaks resolved - which never happens. However, (32) can be used as a preliminary sieve for judging the validity of the relationship

$$\Phi_3(\mathbf{h},\mathbf{k})=\varphi(\mathbf{h})-\varphi(\mathbf{k})-\varphi(\mathbf{h}-\mathbf{k})\approx 0\,(\text{modulo}\,2\pi). \tag{33}$$

A large positive value on the right-hand side of (32) would support the validity of the triple-phase relationship (TPR); a large negative value would make its validity doubtful.

It was said that these were the doldrum years and, in the sense that no critical new developments in phasing were published, this was apparently so. However, during this period a great deal of experience in solving structures was being accumulated - especially by Karle, Hauptman and their co-workers using the results of *ACA Monograph* No. 3 (Hauptman & Karle, 1953). Experience gives understanding and understanding leads to progress. Quietly the bulb had gathered its strength; the flowering was to come.

Progress again: 1963–1969

The method using symbols, suggested by Zachariasen with the modification suggested by Woolfson, became systematized and improved when Karle & Karle (1963) introduced the symbolic addition method for the solution of centrosymmetric structures.

In this technique there is set up a starting set of reflexions with large $|E|$ values, some fixing the origin whose signs can be specified and others which are assigned symbols to represent their signs. We shall illustrate this for the structure of jamine (Karle & Karle, 1964*b*) using material provided by I. L. Karle for a NATO Advanced Study Institute held in York in 1980. The starting set was

h	$s(\mathbf{h})$	$\lvert E(\mathbf{h})\rvert$	
1 $\bar{1}$ 7	+	3·76	origin fixing
2 1 4	+	6·88	origin fixing
3 5 2	+	4·74	origin fixing
0 2 10	*a*	2·63	
3 6 1	*b*	4·82	
0 1 9	*c*	2·28	
1 $\bar{9}$ 2	*d*	3·14	
0 2 $\bar{2}$	*f*	2·30.	

Karle & Karle set up what is now called a $\sum_2$ list [see (23*b*)], that is, a list of pairs of reflexions the product of whose signs give an indication of the sign of a reflexion with a particular vector index $\mathbf{h}$. From these is constructed a table such that from the starting set of signs new sign indications may be found. For jamine the first ten steps in this process are:

(1)	2 1 4	+	
	0 $\bar{2}$ 10	*a*	
	2 $\bar{1}$ $\bar{6}$	*a*	
(2)	2 1 4	+	
	$\bar{1}$ 1 7	±	
	1 2 11	+	
(3)	$\bar{2}$ $\bar{1}$ $\bar{4}$	+	
	3 5 2	±	
	1 4 $\bar{2}$	+	
(4)	1 $\bar{1}$ $\bar{7}$	+	
	3 5 2	±	
	4 4 $\bar{5}$	+	
(5)	1 $\bar{1}$ $\bar{7}$	+	
	1 4 $\bar{2}$	±	
	2 3 $\bar{9}$	+	
(6)	3 5 2	+	
	0 $\bar{2}$ $\bar{10}$	*a*	
	3 3 $\bar{8}$	*a*	
	2 $\bar{1}$ $\bar{6}$	*a*	
	1 4 $\bar{2}$	±	
	3 3 $\bar{8}$	*a*	
(7)	1 4 $\bar{2}$	+	
	0 $\bar{2}$ $\bar{10}$	*a*	
	1 2 $\bar{12}$	*a*	

(8)
$$\begin{array}{cccc} 1 & \bar{1} & \bar{7} & + \\ \underline{0} & \underline{2} & \underline{10} & \underline{a} \\ 1 & 1 & 3 & a \end{array}$$
$$\begin{array}{cccc} 4 & 4 & \bar{5} & + \\ \underline{\bar{3}} & \underline{\bar{3}} & \underline{8} & \underline{a} \\ 1 & 1 & 3 & a \end{array}$$
$$\begin{array}{cccc} 2 & 3 & \bar{9} & + \\ \underline{\bar{1}} & \underline{2} & \underline{12} & \underline{a} \\ 1 & 1 & 3 & a \end{array}$$

(9)
$$\begin{array}{cccc} 3 & 6 & 1 & b \\ \underline{1} & \underline{\bar{1}} & \underline{\bar{7}} & \underline{+} \\ 4 & 5 & \bar{6} & b \end{array}$$

(10)
$$\begin{array}{cccc} \bar{2} & \bar{1} & \bar{4} & + \\ \underline{4} & \underline{5} & \underline{\bar{6}} & \underline{b} \\ 2 & 4 & \overline{10} & b \end{array}$$
$$\begin{array}{cccc} 3 & 6 & 1 & b \\ \underline{\bar{1}} & \underline{2} & \underline{\overline{11}} & \underline{+} \\ 2 & 4 & \overline{10} & b \end{array}.$$

It can be seen that where there are multiple indications they are consistent and, for this structure, that pattern is maintained for most of the sign-determining process. In step (50) there is found

(50)
$$\begin{array}{cccc} 2 & 1 & 4 & + \\ \underline{1} & \underline{1} & \underline{1} & \underline{bc} \\ 3 & 2 & 5 & bc \end{array}$$
$$\begin{array}{cccc} 2 & \bar{8} & 3 & bcd \\ \underline{1} & \underline{10} & \underline{2} & \underline{d} \\ 3 & 2 & 5 & bc \end{array}$$
$$\begin{array}{cccc} 2 & \bar{4} & \bar{1} & bd \\ \underline{1} & \underline{6} & \underline{6} & \underline{bc} \\ 3 & 2 & 5 & cd \end{array}$$

which indicates that $s(3\,2\,5) \approx bc$ and $b \approx d$. Later in the process, at step (60) there is found

(60)
$$\begin{array}{cccc} \bar{2} & 0 & 6 & bc \\ \underline{2} & \underline{4} & \underline{\overline{10}} & \underline{b} \\ 0 & 4 & \bar{4} & c \end{array}$$
$$\begin{array}{cccc} \bar{1} & \bar{1} & \bar{1} & bc \\ \underline{1} & \underline{5} & \underline{\bar{3}} & \underline{b} \\ 0 & 4 & \bar{4} & c \end{array}$$
$$\begin{array}{cccc} 1 & 10 & 2 & d \\ \underline{\bar{1}} & \underline{\bar{6}} & \underline{\bar{6}} & \underline{bc} \\ 0 & 4 & \bar{4} & bcd \end{array}$$
$$\begin{array}{cccc} \bar{2} & 8 & \bar{3} & bcd \\ \underline{2} & \underline{\bar{4}} & \underline{\bar{1}} & \underline{bd} \\ 0 & 4 & \bar{4} & c \end{array}$$
$$\begin{array}{cccc} 2 & \bar{5} & \overline{10} & bcd \\ \underline{\bar{2}} & \underline{9} & \underline{6} & \underline{bd} \\ 0 & 4 & \bar{4} & c \end{array}$$
$$\begin{array}{cccc} 1 & 6 & 7 & c \\ \underline{\bar{1}} & \underline{2} & \underline{\overline{11}} & \underline{+} \\ 0 & 4 & \bar{4} & c \end{array}$$

where, once again, there is an indication $b \approx d$. Actually it turned out that $b = d$ was incorrect and I. L. Karle used this example to illustrate that great caution must be exercised in accepting such indications.

Some possible combinations of signs for *abcdf* yield too many failures of sign relationships and may be rejected. Sometimes it is possible to use the $\sum_1$ formula [(23*a*)] to indicate one or more of the signs and $c = +$ was accepted on this basis. For the sixteen (2^4) possible sets of signs for *abdf* eight gave rather indeterminate sign indications for 34 or more of the reflexions being phased and so were eliminated. The solution with all symbols positive could also be eliminated, because this would lead to all positive signs, and another set with a large predominance of positive signs could also be discounted. The *E* map which showed the structure is shown in Fig. 4.

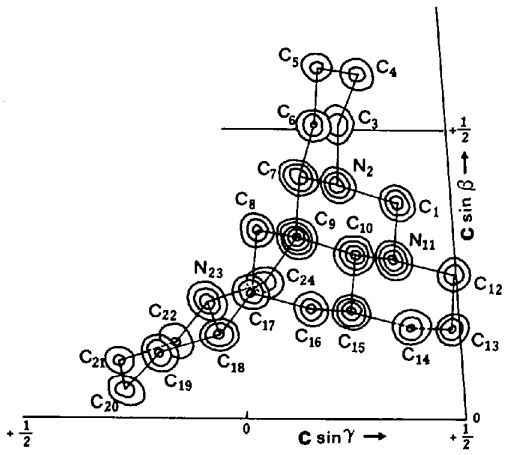

Fig. 4. Sections of *E* map for jamine calculated with 286 terms for which $|E| > 1{\cdot}36$.

The real importance of this systematic approach to the application of symbols became evident when Karle & Karle (1964*a*) used a symbolic addition procedure in the first direct-methods solution of a non-centrosymmetric structure, L-arginine dihydrate. It is difficult to overstate the importance of this step. Many had thought that the problem of determining general phases, which could be anywhere in the range 0 to 2π, would be impossibly difficult and few were inclined even to try to do so. However, once it was demonstrated that general phases could be found then a great deal of effort was devoted to finding ever better processes for doing so. We illustrate the use of symbolic addition with material provided by I. L. Karle for the NATO Advanced Study Institute held in York in 1980.

The starting set for L-arginine dihydrate, space group $P2_12_12_1$, was

h	$\lvert E(\mathbf{h})\rvert$	$\varphi(\mathbf{h})$	
2 0 10	3·46	0	origin specification
3 3 0	2·17	$-\pi/2$	origin specification
3 0 1	2·77	$\pi/2$	origin specification
2 12 0	3·21	P	must be 0 or π
2 10 0	2·31	s	must be 0 or π
4 0 14	2·56	m	must be 0 or π
3 8 3	2·31	a.	

For a non-centrosymmetric structure it is possible to specify not only an origin but also an enantiomorph. However, Karle & Karle declined to take the opportunity of fixing an extra phase in the starting set and they specified the enantiomorph later in the process.

It is instructive to look at a selection of the early steps of the development of phases because this illustrates the features which arise in symbolic addition with non-centrosymmetric structures.

(1)
$$\begin{array}{ccc} 3 & 0 & 1 \\ \underline{3} & \underline{3} & \underline{0} \\ 6 & 3 & 1 \end{array} \qquad \begin{array}{c} \pi/2 \\ \underline{-\pi/2} \\ 0 \end{array}$$

(2)
$$\begin{array}{ccc} \bar{3} & 0 & 10 \\ \underline{6} & \underline{3} & \underline{1} \\ 3 & 3 & 11 \end{array} \qquad \begin{array}{c} \pi \\ \underline{0} \\ \pi \end{array}$$

(3)
$$\begin{array}{ccc} \bar{4} & 0 & 14 \\ \underline{6} & \underline{3} & \underline{\bar{1}} \\ 2 & 3 & 13 \end{array} \qquad \begin{array}{c} m \\ \underline{\pi} \\ \pi+m \end{array}$$

(4)
$$\begin{array}{ccc} \bar{3} & 0 & \bar{1} \\ \underline{3} & \underline{3} & \underline{11} \\ 0 & 3 & 10 \end{array} \qquad \begin{array}{c} -\pi/2 \\ \underline{\pi} \\ \pi/2 \end{array}$$

$$\begin{array}{ccc} 3 & 3 & 0 \\ \underline{\bar{3}} & \underline{0} & \underline{10} \\ 0 & 3 & 10 \end{array} \qquad \begin{array}{c} -\pi/2 \\ \underline{\pi} \\ \pi/2 \end{array}$$

(8)
$$\begin{array}{ccc} 5 & 3 & \bar{3} \\ \underline{\bar{3}} & \underline{\bar{3}} & \underline{11} \\ 2 & 0 & 8 \end{array} \qquad \begin{array}{c} m \\ \underline{\pi} \\ \pi+m \end{array}$$

$$\begin{array}{ccc} 1 & 0 & 4 \\ \underline{1} & \underline{0} & \underline{4} \\ 2 & 0 & 8 \end{array} \qquad \begin{array}{c} m \\ \underline{m} \\ 0 \end{array}$$

Suggests $m = \pi$

(11)
$$\begin{array}{ccc} 2 & 0 & 5 \\ \underline{\bar{1}} & \underline{0} & \underline{4} \\ 1 & 0 & 9 \end{array} \qquad \begin{array}{c} -\pi/2+m \\ \underline{\pi+m} \\ \pi/2 \end{array}$$

$$\begin{array}{ccc} 3 & 0 & 1 \\ \underline{\bar{2}} & \underline{0} & \underline{8} \\ 1 & 0 & 9 \end{array} \qquad \begin{array}{c} \pi/2 \\ \underline{\pi+m} \\ -\pi/2+m \end{array}$$

Suggests $m = \pi$

(12)

3 0 $\bar{10}$	π
1 3 17	0
4 3 7	π
$\bar{1}$ 0 4	$\pi-m$
5 3 3	m
4 3 7	π
1 0 $\bar{4}$	$\pi+m$
3 3 11	π
4 3 7	m
$\bar{2}$ 0 8	$\pi+m$
6 3 $\bar{1}$	π
4 3 7	m

Suggests $m=\pi$

(15)

2 12 0	P
$\bar{1}$ 0 4	0
1 12 4	P

$\bar{1}$ 12 $\bar{1}$	$3\pi/2+P$
2 0 5	$\pi/2$
1 12 4	P
3 12 9	$\pi/2+P$
$\bar{2}$ 0 $\bar{5}$	$-\pi/2$
1 12 4	P

(18)

3 0 10	0
2 0 $\bar{8}$	0
5 0 2	0
3 0 $\bar{1}$	$-\pi/2$
2 0 3	$-\pi/2$
5 0 2	π
3 $\bar{3}$ $\bar{11}$	π
2 3 13	0
5 0 2	π

First inconsistency.

The specification of the enantiomorph came about from the following steps:

(28)

1 10 10	s
$\bar{1}$ 0 $\bar{4}$	π
0 10 6	$\pi+s$
$\bar{2}$ 10 $\bar{6}$	s
2 0 12	$\pi-2a$
0 10 6	$\pi+s-2a$
2 10 0	s
$\bar{2}$ 0 6	π
0 10 6	$\pi+s$

(31)

$\bar{3}$ 0 10	π
3 8 $\bar{3}$	$-a$
0 8 7	$\pi-a$
3 11 $\bar{4}$	a
$\bar{3}$ $\bar{3}$ 11	π
0 8 7	$\pi+a$.

Steps (28) suggested that $a=0$ or π, which then made the two indications in (31) consistent. If we consider the two origin-fixing reflexions 301 and 3,0,10 then $\varphi(3\,0\,1)+\varphi(3\,0\,10)+\varphi(0\,8\,7)$ forms a structure semi-invariant whose value takes one of the values $\pm\pi/2$. By making $a=0$ then the semi-invariant value is fixed at $\pi/2$ and the enantiomorph is fixed.

It was possible to find probable values for many of the symbols from the $\sum_1$ formula which for this space group gives

$$s(2h\,0\,2l)\approx s\left[\sum_k(-1)^{k+l}\{|E(h\,k\,l)|^2-1\}\right]=s(\sum\nolimits_1) \tag{34}$$

with similar expressions for $s(0\,2k\,2l)$ and $s(2h\,2k\,0)$. There is an associated probability

$$P_+[(2\mathbf{h})]=\tfrac{1}{2}+\tfrac{1}{2}\tanh\,[\sigma_3|E(2\mathbf{h})|\textstyle\sum_1/2\sigma_2^{3/2}]. \tag{35}$$

Thus one could find

2h	Symbol	$P_+[E(2\mathbf{h})]$
2 12 0	P	0·14
4 8 0	$\pi+p$	0·72
0 12 8	P	0·27

and a combination of these three probabilities, considered as independent, gives p as π with an overall probability of 0·98.

With estimates for a hundred or so phases a process of phase refinement and extension was embarked upon using the tangent formula, (31). The final E map for L-arginine dihydrate, obtained with estimates for 400 phases with $|E|_{\min}=1{\cdot}00$, is shown in Fig. 5.

This first triumph was followed by many other impressive structure solutions of non-centrosymmetric structures from the use of symbolic addition in the laboratory of Karle & Karle and in other laboratories as well. Karle & Karle relied on the hand application of symbolic addition - although they used a fairly small computer to generate the $\sum_2$ relationships and the phase or sign-development tables.

It was not long before attempts were made to automate completely the whole process and one of the more successful and elaborate programs, although only applicable to centrosymmetric structures, was *LSAM* (logical symbolic addition method) devised by Germain & Woolfson (1968). This combined elements of the ideas introduced by Cochran & Douglas and by Karle & Karle. A starting set is found with origin-fixing reflexions and up to six others to which could be attached sign symbols. The most probable new sign indication was found using the $\sum_2$ relationship, (23*b*); this was added to the starting set and the next most probable sign indication was found. Sometimes a relationship would be found between symbols and these were accumulated; for example with six letter symbols up to 127 relationships were possible. A selection of these for a particular structure with space group $C2/c$ were (with probabilities in parentheses)

$Z_1=-DF\approx+1$ (1·000) $\quad Z_4=AD\approx+1$ (0·985)

$Z_2=ACEF\approx+1$ (1·000) $\quad Z_5=AEF\approx+1$ (0·966)

$Z_3=-EF\approx+1$ (0·999) $\quad Z_6=-AB\approx+1$ (0·944)

and these can be inverted to give

$$A=-Z_3Z_5 \quad B=Z_3Z_5Z_6 \quad C=Z_2Z_5$$

$$D=-Z_3Z_4Z_5 \quad E=-Z_1Z_4Z_5 \quad F=Z_1Z_3Z_4Z_5.$$

LSAM considers possibilities with up to one, or sometimes two, failures in the set Z_1 to Z_6 and then

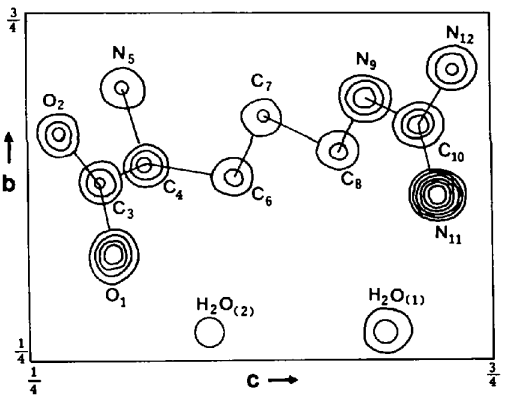

Fig. 5. Sections of E map for L-arginine dihydrate calculated with 400 terms for which $|E|>1{\cdot}00$.

with the resultant absolute values of the letter symbols develops a complete set of signs for several hundred reflexions. Several FOM's were calculated and the user could ask for an E map to be calculated for any selected set of signs. With the aid of the third-generation computers available at the time *LSAM* could solve a complete structure, including production of the E map, in 3-4 min. It is now little used, but it was a trailblazer for better things to come.

During this period Hauptman began an extended program of deriving formulae for the estimation of invariant and semi-invariant quantities. One formula suggested (Hauptman, Fisher, Hancock & Norton, 1969), with a family resemblance to (32), is, for the equal-atom case,

$$\begin{aligned}|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|\cos[\varphi(\mathbf{h})-\varphi(\mathbf{k})-\varphi(\mathbf{h}-\mathbf{k})]\\ \approx K\langle(|E_{\mathbf{H}}|^{1/2}-\overline{|E|^{1/2}})(|E_{\mathbf{h}+\mathbf{H}}|^{1/2}-\overline{|E|^{1/2}})(|E_{\mathbf{k}+\mathbf{H}}|^{1/2}\\ -\overline{|E|^{1/2}})\rangle_{\mathbf{H}}\\ +4N^{-1/2}\{\tfrac{3}{2}[|E(\mathbf{h})E(\mathbf{k})|^2+|E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|^2\\ +|E(\mathbf{h}+\mathbf{k})E(\mathbf{h})|^2]+|E(\mathbf{h})|^2+|E(\mathbf{k})|^2\\ +|E(\mathbf{h}+\mathbf{k})|^2-\tfrac{7}{2}\}\end{aligned} \quad (36)$$

where $\overline{|E|^{1/2}}=\langle|E_{\mathbf{H}}|^{1/2}\rangle_{\mathbf{H}}$ and K is a scale factor chosen in such a way that the calculated distribution of cosines agreed as closely as possible with the theoretical distribution.

From the application of this formula one may derive estimates of the cosine of three-phase invariants in the form

$$\langle\cos[\varphi(\mathbf{h})-\varphi(\mathbf{k})-\varphi(\mathbf{h}-\mathbf{k})]\rangle\approx C(\mathbf{h},\mathbf{k}). \quad (37)$$

Hauptman and his colleagues then set up a function

$$\Phi=\sum_{\mathbf{h},\mathbf{k}} W(\mathbf{h},\mathbf{k})\{\cos[\varphi(\mathbf{h})-\varphi(\mathbf{k})-\varphi(\mathbf{h}-\mathbf{k})] - C(\mathbf{h},\mathbf{k})\}^2, \quad (38)$$

where $W(\mathbf{h},\mathbf{k})$ was a weighting function based on an estimate of the reliability of relationships (37). Phases are then estimated, one at a time, in such a way as to keep Φ at a minimum value.

The method worked for the structure of estriol but it soon became clear that other methods, simpler to apply, would work also. These will be our next topic of consideration.

The computer takes over: 1970–1980

The early days of structural crystallography involved heroic efforts to solve a crystal structure. The collection of data was slow, with the use of photographic film on various kinds of camera after which each spot had to be compared by eye with a standard intensity wedge. A single Fourier synthesis, in two dimensions, meant two or three days of additions of columns of numbers. The computer changed all that: not only could the calculations be done more quickly, almost effortlessly, but computer-controlled diffractometers were developed which, once the crystal was set, would automatically collect intensity data. Direct methods, in which a large number of relationships had to be handled, became a natural application for computers.

The earliest *elaborate* program was *LSAM*, but this was soon followed by *MULTAN* (the multiple-tangent-formula method), a program designed to solve non-centrosymmetric problems (Germain, Main & Woolfson, 1970). The basic philosophy of *MULTAN* was decided before the first line of code was ever written. It was to be that, to the greatest extent possible, what went into the computer was to be the minimum information (*i.e.* unit-cell parameters, space group, cell contents, intensities) compatible with defining the problem and solving the structure. Involvement by the user was to be reduced to a minimum, ideally none at all, and what came out of the computer was to be the structure solution in as clear a form as possible. To give the greatest portability of the program the language used was to be Fortran and only a most basic set of instructions was to be used: all special features of the host computer, no matter how attractive, were to be ignored.

The basic idea for *MULTAN* was derived from consideration of the symbolic addition method. The most critical stage in the phase development process was early on when only a few, sometimes one or two, relationships were available. In symbolic addition phase indications can only be combined when the symbolic part of the indications is the same. If not, then a relationship is obtained between symbols which can only later be used to evaluate the symbols. Germain & Woolfson (1968) suggested that this problem could be overcome by using explicit phase values throughout so that different estimates of a new phase can always be combined by use of the tangent formula. The stages in *MULTAN* are:

(*a*) $|E|$'s are calculated from $|F_{\text{obs}}|$ values. At this stage any known information can be used to derive $|E|$'s which are better at revealing structural information than are the conventional $|E|$ values.

(*b*) $\sum_2$ relationships are found for a subset of the largest E's, sufficient in number to solve the structure. At this stage there are also found the right-hand-side contributors for up to 100 small near-zero $|E|$'s. These are to provide a FOM [see (28)]. Estimates of phases of structure semi-invariants from $\sum_1$ relationships are also made at this stage if they are available.

(*c*) A starting set is selected in the form of origin- and enantiomorph-defining reflexions and a small number of others. This is done by a process called *CONVERGENCE* which eliminates reflexions one by one from the complete set of large $|E|$'s until it has converged on a few, which contain those which will define the origin and enantiomorph and also which will develop new phase information quickly and strongly.

(*d*) Values are assigned to unknown phases in the starting set. This was originally done by 'quadrant permutation' where an unknown general phase would be considered with four possible values, $\pm\pi/4$ and $\pm 3\pi/4$. Special phases would be given their pairs of special values, *e.g.* $\pm\pi/2$ or 0, π *etc.*

(*e*) All permutations of phase for starting-set reflexions are used as initial values for tangent-formula phase extension and refinement. However, a weighted tangent formula is used in which the weight associated with each contribution depends on the estimated reliability of the phases it contains. A weighting scheme suggested by Hull & Irwin (1978) is sometimes used; this is particularly effective for structures without translational symmetry and avoids the trivial solution that all phases are zero.

(*f*) FOM's are calculated to rank the phase sets in order of plausibility. One of these is a measure of how well relationships hold overall, one is like the zero check of Cochran & Douglas and the third a residual-type function suggested by Karle & Karle. This uses values of $E(\mathbf{h})_{\text{calc}}$ computed from

$$K|E(\mathbf{h})|^2_{\text{calc}} = |\sum E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|^2$$

where K is chosen so that

$$K\sum_{\mathbf{h}}|E(\mathbf{h})|^2_{\text{calc}} = \sum_{\mathbf{h}}|E(\mathbf{h})|^2_{\text{obs}}.$$

The residual is

$$R_{\text{Karle}} = \frac{\sum_{\mathbf{h}}||E(\mathbf{h})|_{\text{calc}} - |E(\mathbf{h})|_{\text{obs}}|}{\sum_{\mathbf{h}}|E(\mathbf{h})|_{\text{obs}}}. \qquad (39)$$

The three FOM's are united into a single combined figure of merit (CFOM) which would equal 3·0 for a set of phases best for each individual FOM and 0 for a set worst for each FOM.

(*g*) For the highest value of CFOM, or any other set specified by the user, an E map is calculated. The positions of a number of the highest peaks (usually 1·25 × number of atoms to be found) are output on lineprinter paper in a favourable (spread-out) projection together with tables of 'bond' lengths and 'inter-bond' angles for the peaks and suggested chemical interpretations (Main & Hull, 1978).

An example of a *MULTAN* output is shown in Fig. 6. It will be seen that there are some spurious peaks, inevitably so since there are more peaks than atoms, and some atoms may not appear. In an extreme case only a small fragment of the structure may appear and then one has the problem of trying to build from this to the complete structure. Karle (1968) suggested a process for doing this which is incorporated into *MULTAN*. The fragment is used to compute partial structure factors, $F(\mathbf{h})_p$, and the phase it indicates is accepted if

$$|F(\mathbf{h})_p| > p|F_{\text{obs}}|. \qquad (40)$$

The chosen value of p is governed by a set of rules but is usually in the range 0·25 to 0·6. Accepted phases are then put into the tangent formula and phase information is extended and refined. The following E map, if it does not show most of the structure, should show more of it and the process is repeated until conventional refinement processes can take over.

Over the years *MULTAN* has become a generic name for a variety of methods for generating phases all incorporated in a single package. Two of these, developed in the period under consideration, are worthy of mention. White & Woolfson (1975) introduced a concept which they called 'magic integers'. This states that if there are n phases (in cycles) with values φ_i ($i = 1$ to n) then one can select a set of integers m_i ($i = 1$ to n) such that

$$\varphi_i \simeq m_i x \pmod 1$$

for some value of x in the range 0 to 1.

As an example, by use of three variables, x, y, z, 12 phases may be represented by

$$\begin{pmatrix} 8 \\ 12 \\ 14 \\ 15 \end{pmatrix} (x\,y\,z).$$

Since phases can be represented in magic-integer form, so can the linear combinations of them forming TPR's, and a typical TPR will be in the form (in cycles)

$$\Phi_{3,r} = H_r x + K_r y + L_r z + b_r, \qquad (41)$$

where the b_r results from specifying phases corresponding to reflexions in one asymmetric unit of

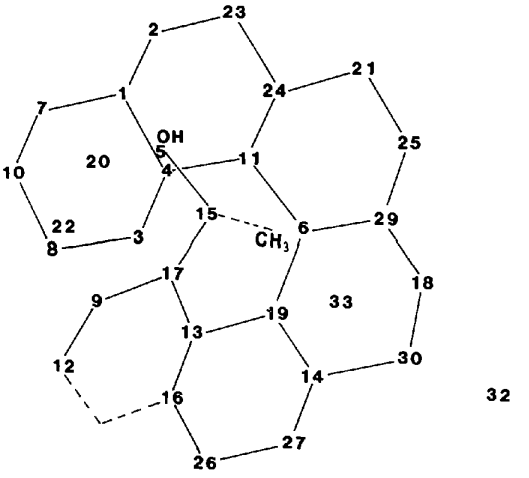

Fig. 6. Representation of a typical *MULTAN* output. The dashed lines are bonds to missing peaks while some spurious peaks are also seen.

reciprocal space for some higher space groups. Since the values of Φ_3 are expected to be close to 0 (mod 1) then one can design a function

$$\psi(x, y, x) = \sum_r \kappa_r[\cos 2\pi(H_r x + K_r y + L_r z + b_r)] \quad (42)$$

which should be large. The right-hand side can be summed as a Fourier map and peaks in it give values of (x, y, z) and hence plausible values for the magic-integer-represented phases.

This principle has been extended in a number of papers and Declercq, Germain & Woolfson (1975) showed how it could be used effectively to get a starting set of 50–60 phases for tangent-formula extension. Many structures for which initial runs of *MULTAN* failed have succumbed to a single run of the program *MAGIC*.

The magic-integer principle solves the problem of how to populate a many-dimensional phase space as uniformly as possible with a given number of points. A complete theory has been worked out by Main (1977). In fact, from 1980 phase permutation, described as stage (d) in the *MULTAN* description, has been replaced by magic-integer permutation.

The second development incorporated into the *MULTAN* package had a rather serendipitous origin. Woolfson (1977) proposed an idea for refining phases by expressing TPR's as linear equations so that, with phases in cycles, a TPR was of the form

$$\varphi_r \pm \varphi_s \pm \varphi_t + b \approx n, \text{ an integer.}$$

The value of n could be estimated as the nearest integer to the left-hand side with current phase values substituted and a complete set of equations appears as

$$\mathbf{A}\boldsymbol{\varphi} + \mathbf{b} \approx \mathbf{n}. \quad (43)$$

A least-squares solution for $\boldsymbol{\varphi}$ gives

$$\boldsymbol{\varphi} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{n} - \mathbf{b}), \quad (44)$$

which gives new phase estimates for another cycle of refinement. Early trials showed that the phases so obtained had less error than those derived from the tangent formula.

For any refinement process it is rational to try to find its radius of convergence. A systematic investigation of this, in which the initial phase errors were gradually increased, gave the surprising result that in a significant proportion of trials initially random phases would refine to correct values.

It turns out that starting with random phases has some advantages over the conventional *MULTAN* approach. If a convergence map is followed by symbolic addition it is often found that the first few steps give completely consistent, and sometimes wrong, phase indications so that no matter what are the starting phases in *MULTAN* the process will go wrong. If a start is made with a large number of random phases then so many phase relationships are being deployed simultaneously that the idiosyncratic behaviour of a few of them does little harm. Some versions of *MULTAN* contain a program *YZARC* which exploits the idea of starting with random phases and refining with linear equations. This too is often successful when *MULTAN* fails.

In parallel with the development of methods of applying TPR's, a great deal of work was in progress to find formulae from which the values of phase-dependent structure invariants could be estimated. The leading figure in this work was Hauptman (1975) who put forward what he called 'the neighbourhood principle'. This stated that, for a particular enantiomorph, the value of a structure invariant could be estimated by the values of one or more sets of few magnitudes $|E|$, the neighbourhoods of the structure invariant. As an example of the application of this principle we consider the structure invariant known as a quartet, which has had some applications in practical direct methods. This is

$$\Phi_4 = \varphi(\mathbf{h}) + \varphi(\mathbf{k}) + \varphi(\mathbf{l}) + \varphi(\mathbf{m}) \quad (45)$$

where

$$\mathbf{h} + \mathbf{k} + \mathbf{l} + \mathbf{m} = \mathbf{0}.$$

If $|E(\mathbf{h})|$, $|E(\mathbf{k})|$, $|E(\mathbf{l})|$ and $|E(\mathbf{m})|$ are all large then it can be shown that the probability distribution of Φ_4 resembles that of Φ_3, shown in Fig. 3. However, on the basis of the neighbourhood concept there are three other magnitudes which strongly influence the value of Φ_4; these are

$$|E(\mathbf{h}+\mathbf{k})|, \quad |E(\mathbf{h}+\mathbf{l})| \quad \text{and} \quad |E(\mathbf{h}+\mathbf{m})|,$$

usually referred to as the 'cross terms'. Hauptman has shown that if the cross terms are all large then the conclusion that $\Phi_4 \approx 0$ (modulo 2π) is strongly reinforced - a probability density curve would be found, as in Fig. 7(a), which had a much smaller variance than that found just from the four main magnitudes. An interesting case arises when the cross terms are all small. The distribution now appears as in Fig. 7(b) and leads to the conclusion that $\Phi_4 \approx \pi$ (modulo 2π); such a combination of four phases is called a 'negative quartet'. It is proper to mention here that the concept that the magnitudes of the three cross terms controlled the probable value of a quartet originated with Schenk (1973a), Schenk & de Jong (1973) and Schenk (1973b). Intermediate cases can also arise where, for example, two of the cross terms are large and one small, when a bimodal distribution as shown in Fig. 7(c) may result.

The neighbourhood principle has been applied with more magnitudes (*e.g.* 15) to estimate quartets and also to estimate triples (TPR's), quintets, sextets *etc.*, but these latter have more curiosity than practical value - at least at present.

Work along similar lines to estimate the values of structure invariants and semi-invariants has been pursued by Giacovazzo (1977), using what he calls the 'theory of representations'. This leads to a concept of 'phasing shells' which are similar to, but not the same as, the neighbourhoods of Hauptman. There has been some debate about the relative merit of the two approaches; it is probably fair to say that the phasing shells give a more complete selection of relevant magnitudes but that the neighbourhood principle picks out of them those that are most important. The principle of neighbourhoods or representation theory is included implicitly in a general theory of invariants and embedded semi-invariants given by Karle (1982*a*).

Quartets have been used in various ways in direct methods. A four-phase invariant may be used to estimate one phase when three others are known and they may thus be incorporated into the phasing process of, say, *MULTAN*. This has been done by Gilmore (1984) in the program *MITHRIL*, which contains *MULTAN* as its main component but includes other methods as well, by Sheldrick (1975) in the program *SHELX* which also incorporates the *MULTAN* approach, and by Overbeek & Schenk (1978) in *SIMPEL*, an automated symbolic addition programme. The quartet contributions to phasing have usually been incorporated within a modified tangent formula of the form

$$\tan[\varphi(\mathbf{h})] = (B+D)/(A+C) \qquad (46)$$

where

$$B = \sum_{\mathbf{k}} |E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \sin[\varphi(\mathbf{k})+\varphi(\mathbf{h}-\mathbf{k})] \qquad (47a)$$

$$A = \sum_{\mathbf{k}} |E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \cos[\varphi(\mathbf{k})+\varphi(\mathbf{h}-\mathbf{k})] \qquad (47b)$$

$$D = n \sum_{\mathbf{l}} \sum_{\mathbf{m}} |E(\mathbf{l})E(\mathbf{m})E(\mathbf{h}-\mathbf{l}-\mathbf{m})| \times \sin[\varphi(\mathbf{l})+\varphi(\mathbf{m})+\varphi(\mathbf{h}-\mathbf{l}-\mathbf{m})] \qquad (47c)$$

and

$$C = n \sum_{\mathbf{l}} \sum_{\mathbf{m}} |E(\mathbf{l})E(\mathbf{m})E(\mathbf{h}-\mathbf{l}-\mathbf{m})| \times \cos[\varphi(\mathbf{l})+\varphi(\mathbf{m})+\varphi(\mathbf{h}-\mathbf{l}-\mathbf{m})], \qquad (47d)$$

where n is a positive constant chosen either empirically or from a theoretical base depending on the probability distributions of the triplets and quartets.

By and large, the most useful application of the quartets has been of negative quartets, the sum of four phases with a probability density peaked at π. In the space groups $P1$, $P\bar{1}$, and other symmorphic space groups, these are a way of introducing negative relationships between signs or non-zero differences between phases where, otherwise, the trivial solution $\varphi = 0$ (or $s = +1$) for all phases (signs) satisfies all three phase (sign) relationships. Negative quartets have also been found to give a useful FOM. De Titta, Edmonds, Langs & Hauptman (1975) proposed the FOM

$$\text{NQEST} = \sum_i |E^4|_i \cos \Phi_{4,i} \Big/ \sum_i |E^4|_i, \qquad (48)$$

where the $\Phi_{4,i}$ are negative quartets and $|E^4|_i$ is the appropriate product of four strong $|E|$'s. The summation is made over as many strong negative quartets as are available (usually up to 99) and the most negative value of NQEST usually indicates the correct solution.

It was in this period that the first applications of direct methods were made to protein structures. The great triumphs of protein crystallography had resulted mostly from the application of the multiple-isomorphous-replacement (MIR) method. By comparison of the intensities from a native protein and several isomers, usually produced by adding heavy-atom-containing groups to various points of the protein chain, estimates of phases can be made. These are usually rather imprecise (r.m.s. error 40–50°) and

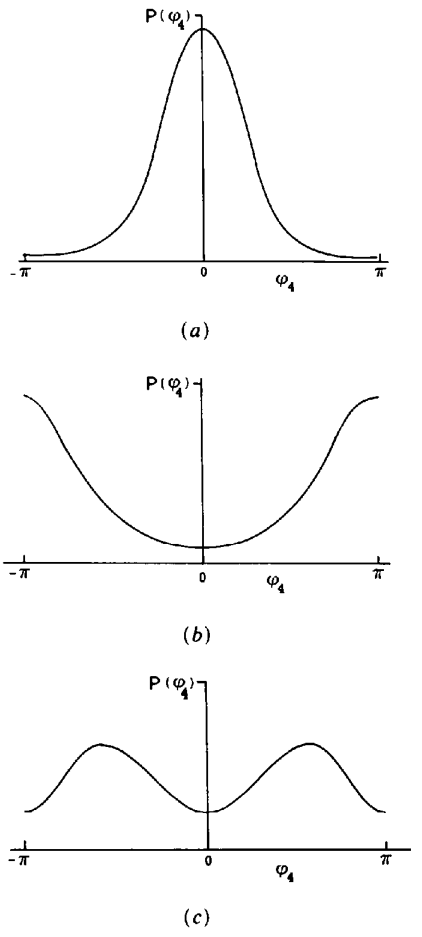

Fig. 7. Possible distribution functions for a phase quartet with (*a*) all cross terms large, (*b*) all cross terms small, and (*c*) cross terms not all large or all small.

restricted to low-resolution reflexions because of the limited extent of isomorphism of the compounds. Techniques for phase extension and refinement are required both to improve the knowledge of low-resolution phases and also to extend phase knowledge to higher resolution. Most early methods involved density modification - *i.e.* production of an electron density map, modification of it to conform to some required property and then transformation of the map to estimate new phases.

An extension of the use of Karle–Hauptman determinants [(14)] was proposed by Tsoucaris (1970) with his maximum determinant rule. One property of the Karle–Hauptman determinants which was given by Goedkoop (1950) and Hauptman & Karle (1950*a*) is that when the order of the determinant is greater than the number of atoms in the unit cell, N, then its value equals zero. For a determinant of order $m \leq N$ the value is positive and, in addition, all the eigenvalues of the Karle–Hauptman matrix must be positive.

With this background we can now state the maximum determinant rule: Assume that all the structure factors are known, in magnitude and phase in D_m, a determinant of order m. If a new determinant, D_{m+1}, is constructed by adding a new row and column the most probable phases of the $(m+1)$th row (or column) are those which lead to the maximum value of D_{m+1} under the condition that the matrix eigenvalues are all positive.

This rule has been used to extend phases in protein crystallography (Mauguen, 1979) by the use of what is known as the 'regression equation' which expresses the most probable value of one structure factor as a function of all the other structure factors in D_{m+1} (de Rango, Tsoucaris & Zelwer, 1974). With this iterative procedure a maximization of D_{m+1} is possible by gradual modification of the phases in the final column.

Another pioneering venture in phase extension and refinement was carried out in this period by Sayre (1972, 1974). This was done by the use of the Sayre equation (18) and an iterative full-matrix least-squares process for deriving phases which minimizes the function

$$S\langle\varphi\rangle = \sum_{\mathbf{h}} w(\mathbf{h})|F(\mathbf{h}) - \theta(\mathbf{h}) \sum_{\mathbf{k}} F(\mathbf{k})F(\mathbf{h}-\mathbf{k})|^2, \quad (49)$$

where $w(\mathbf{h})$ is a weight. The method was effective but very costly in computer time. For rubredoxin and insulin (Cutfield, Dodson, Dodson, Hodgkin, Isaacs, Sakabe & Sakabe, 1975), 2·5 and 1·9 Å heavy-atom phases were used as input data and the phases were extended to 1·5 Å to give readily interpretable maps.

The Sayre method represents a very effective way of using sheer brute computer power. However, such work can only be done in a few places, well favoured with computer resources.

A development of this period, which may become important as massive computer power becomes available, is the development of a generalized tangent formula by Karle (1971). This is based on the determinantal inequalities and is of the form

$$\varphi(\mathbf{h}) = \text{phase of } \left\{\sum_P \delta_{m,p}(\mathbf{h})\right\}, \quad (50)$$

where $\delta_{m,p}(\mathbf{h})$ is a ratio of two order $m-1$ determinants derived from the basic order m Karle–Hauptman determinant,

$$D_{m,p}(\mathbf{h}) = \begin{vmatrix} E(\mathbf{0}) & E(\bar{\mathbf{k}}_1) & \cdots & E(\bar{\mathbf{k}}_{m-2}) & E(\bar{\mathbf{h}}) \\ E(\mathbf{k}_1) & E(\mathbf{0}) & \cdots & E(\mathbf{k}_1-\mathbf{k}_{m-2}) & E(\mathbf{k}_1-\mathbf{h}) \\ E(\mathbf{k}_2) & E(\mathbf{h}_2-\mathbf{k}_1) & \cdots & E(\mathbf{k}_2-\mathbf{k}_{m-2}) & E(\mathbf{k}_2-\mathbf{h}) \\ \vdots & \vdots & & \vdots & \vdots \\ E(\mathbf{k}_{m-2}) & E(\mathbf{k}_{m-2}-\mathbf{k}_1) & \cdots & E(\mathbf{0}) & E(\mathbf{k}_{m-2}-\mathbf{h}) \\ E(\mathbf{h}) & E(\mathbf{h}-\mathbf{k}_1) & \cdots & E(\mathbf{h}-\mathbf{k}_{m-2}) & E(\mathbf{0}) \end{vmatrix} \quad (51)$$

and p indicates a particular set of vectors $\mathbf{k}_1$ to $\mathbf{k}_{m-2}$.

For $m = 3$ the result is the normal tangent formula (31). Larger values of m yield formulae giving better estimates of $\varphi(\mathbf{h})$ although at the expense of increased computation. Indeed, for a high enough value of m the phase, $\varphi(\mathbf{h})$, may be given precisely.

Modern times: 1981–1986

The early years of the 1980's saw the extension of earlier work in developing methods of solving small structures. Programs such as *SIMPEL* and *SHELX* were extended and made more versatile and there was also a move in the direction of making these programs, and *MULTAN*, available on minicomputers and even microcomputers, and in extending their use to structures with abnormal features, *e.g.* those with pseudo-translational symmetry (Fan Hai-fu, Yao Jia-xing, Main & Woolfson, 1983). However, there were also some interesting new concepts introduced which extended either the range of applicability of direct methods or at least the efficiency of their application.

An interesting development of the 'random' approach was made by Yao Jia-xing (1981) in the development of *RANTAN* - random *MULTAN*. The advantages of the *MAGIC* and *YZARC* approaches in increasing the size of the starting set of reflexions has already been mentioned. Yao took this to its logical limit by doing without a starting set - or, perhaps more precisely, taking all the structure factors whose phases were needed as the starting set. In this multisolution method, for each trial random phases are allocated to all structure factors, other than those picked to fix the origin and enantiomorph, and these phases are then refined by a very controlled use of the weighted-tangent formula. The effectiveness of this method was demonstrated for many structures -

for example complex 2 enniantin C:1 KSCN, with space group $P2_1$ and 100 atoms in the asymmetric unit (Yao Jia-xing, 1983). After making 78 trials *RANTAN* stopped automatically, since it found good figures of merit, and the resulting E map showed 85 of the 100 atoms to be found. Yao also tried his method on a synthetic structure in space group $P1$ containing three molecules of valinomycin, giving 234 atoms to be found. Although he degraded the calculated structure factors to simulate the quality of observed data he was still able to solve the structure, finding 192 atoms in the first E map he examined. The *RANTAN* principle is so effective that it has been incorporated as the default method in the *MULTAN* system.

In his 1983 paper Yao also demonstrated how the *RANTAN* principle could be used as a powerful method of fragment development by a multisolution approach. A number of phases were accepted by means of the Karle criterion given in relationship (40) and to all remaining reflexions, whose phases did not satisfy the relationship, random values were assigned. These phases were then taken through the *RANTAN* process. Usually, in comparatively few trials, a set of phases with good FOMs would be found showing most of the structure. One example of this process was with the known structure of virginiamycin factor-S methanol solvate (Declercq, Germain, Van Meerssche, Hull & Irwin, 1978) with 66 independent atoms in space group $P2_12_12_1$. Neither *MULTAN* nor *RANTAN* could solve it in a straightforward way but by starting with the positions of 12 atoms the structure could be obtained by Karle recycling, successive cycles giving 16, 26, 42, 66 atoms. With the *RANTAN* method, starting with only six known atom positions, the fifth trial gave good FOMs and showed 52 atoms.

Another approach of a rather general kind has been described by Debaerdemaeker & Woolfson (1983). They considered a number of functions of the phases, which would be expected to be either a maximum or minimum, and then from a starting point of random phases they refined to the required extremum by a parameter-shift process. This involves changing the phases, one at a time, over a range of values and taking the best value in the range. In fact this is what is done by the tangent formula. The function being maximized is

$$Q = \sum_{\mathbf{h}} \sum_{\mathbf{k}} \kappa(\mathbf{h}, \mathbf{k}) \cos [\varphi(\mathbf{h}) - \varphi(\mathbf{k}) - \varphi(\mathbf{h}-\mathbf{k})]. \quad (52)$$

A maximum requires $\partial Q/\partial\varphi(\mathbf{h}) = 0$ for all $\mathbf{h}$ and application of this condition gives

$$\sum_{\mathbf{k}} \kappa(\mathbf{h}, \mathbf{k}) \sin [\varphi(\mathbf{h}) - \varphi(\mathbf{k}) - \varphi(\mathbf{h}-\mathbf{k})] = 0. \quad (53)$$

A rearrangement of this equation gives the tangent formula, (31). The value of $\varphi(\mathbf{h})$ so found gives the maximum value of Q, given all the current values of other phases - which is the basis of the parameter-shift approach.

An example of a function to be maximized, explored by Debaerdemaeker & Woolfson (1983), is

$$\psi_C = \sum_{\mathbf{h}} [X(\mathbf{h}) - |Y(\mathbf{h})|] \quad (54)$$

where

$$X(\mathbf{h}) = \sum_{\mathbf{k}} \kappa(\mathbf{h}, \mathbf{k}) \cos [\varphi(\mathbf{h}) - \varphi(\mathbf{k}) - \varphi(\mathbf{h}-\mathbf{k})] \quad (55a)$$

and

$$Y(\mathbf{h}) = \sum_{\mathbf{k}} \kappa(\mathbf{h}, \mathbf{k}) \sin [\varphi(\mathbf{h}) - \varphi(\mathbf{k}) - \varphi(\mathbf{h}-\mathbf{k})]. \quad (55b)$$

Maximization will tend to give the greatest possible value of the summation over $X(\mathbf{h})$, which is equivalent to Q in (50), while having the smallest possible value for the summation over $|Y(\mathbf{h})|$. By chance, owing to a mistake in the computer programming a much better and very powerful function to be maximized was found, namely

$$\psi_D = \sum [X(\mathbf{h}) - Y(\mathbf{h})]. \quad (56)$$

There is no rational basis for using this function but, nevertheless, maximizing ψ_D vies in power with any other method found and has accounted for numerous successful crystal-structure solutions with up to 200 independent atoms. If the $X - Y$ method has been described in some detail it is to underline the fact that serendipity has as powerful a role to play as the intellect in this as in many other branches of science.

The majority of the methods described so far depend on probabilistic relationships between phases, but it might be expected that a phase-determining procedure based on an exact relationship would be stronger. Debaerdemaeker, Tate & Woolfson (1985) devised a new tangent formula based on satisfying the Sayre equation. Following the philosophy by which the normal tangent formula could be derived by maximizing Q, in (52), they looked at the minimization of

$$R = \sum_{\mathbf{h}} |E(\mathbf{h}) - KG(\mathbf{h})|^2 \Big/ \sum_{\mathbf{h}} |E(\mathbf{h})|^2, \quad (57)$$

where

$$G(\mathbf{h}) = [1/g(\mathbf{h})] \sum_{\mathbf{k}} E(\mathbf{k}) E(\mathbf{h}-\mathbf{k}) \quad (58)$$

and $g(\mathbf{h})$ is known. If a value of K can be found which makes $R = 0$ then Sayre's equation would be exactly obeyed for the data set. The value of K could be found from

$$\partial R/\partial K = 0, \quad (59)$$

and then phases had to satisfy the condition

$$\partial R/\partial\varphi(\mathbf{h})=0 \text{ for all } \mathbf{h}. \tag{60}$$

This led to the result

$$\tan[\varphi(\mathbf{l})]=\frac{\text{Im}[t(\mathbf{l})]-(2T/3Q)\,\text{Im}[q(\mathbf{l})]}{\text{Re}[t(\mathbf{l})]-(2T/3Q)\,\text{Re}[q(\mathbf{l})]} \tag{61}$$

where

$$t(\mathbf{l})=\sum_{\mathbf{h}}[1/g(\mathbf{l})+1/g(\mathbf{h})+1/g(\mathbf{l}-\mathbf{h})]|E(\mathbf{h})E(\mathbf{l}-\mathbf{h})| \times\exp\{i[\varphi(\mathbf{h})+\varphi(\mathbf{l}-\mathbf{h})]\} \tag{62}$$

$$q(\mathbf{l})=\sum_{\mathbf{h}}|E(\mathbf{l}-\mathbf{h})|\exp[i\varphi(\mathbf{l}-\mathbf{h})][1/g(\mathbf{h})^2] \times\sum_{\mathbf{k}}|E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|\exp\{i[\varphi(\mathbf{k})+\varphi(\mathbf{h}-\mathbf{k})]\} \tag{63}$$

$$T=\sum_{\mathbf{l}}E(\mathbf{l})^*t(\mathbf{l}) \tag{64}$$

and

$$Q=\sum_{\mathbf{l}}E(\mathbf{l})^*q(\mathbf{l})=\sum_{\mathbf{h}}|G(\mathbf{h})|^2. \tag{65}$$

The terms in $q(\mathbf{l})$ come from a special set of quartet relationships, those with $E(\mathbf{l})$, $E(\mathbf{l}-\mathbf{h})$, $E(\mathbf{k})$ and $E(\mathbf{h}-\mathbf{k})$ belonging to the set of large E's but with the cross terms $E(\mathbf{h})$ belonging either to the set of large E's or to a selected set of small (ideally zero) E's. The net effect of the refinement is to produce a set of phases for the large E's which satisfy Sayre's equation both for large E's and small E's. This is the first example of the explicit routine and large-scale use of small-E data to determine phases - although their use might be implicit in some determinantal methods with small E's as elements, and small E's had been used in the $\sum_3$ formula [(23c)] in some early work. Despite its apparent complexity the Sayre tangent formula, (59), is simple to apply and takes only one and a half times as long to refine a set of phases as does the normal tangent formula. The mode of use is to use trial sets of random phases, as in *RANTAN*, and the final phase sets are judged by the same figures of merit as are used in the conventional *MULTAN* procedure. The Sayre tangent formula method, incorporated as *SAYTAN* in the most recent *MULTAN* package, is considerably more powerful than those using the conventional tangent formula with a power only matched by the enigmatic $X-Y$ technique.

The combination of triplet and quartet terms which appear in the Sayre tangent formula also occurs in the modified tangent formula (46) although with the *sign* of the quartet term reversed. The generalized tangent formula (50), with $m=4$, also gives triplet and quartet terms, in this case with different weights from those indicated in (61). Although the generalized tangent formula, in the form given, has never actually been applied it is possible that it would give similar results to formula (61), although it may be difficult systematically to choose terms in the summation in (50) so as to include a particular set of large and small $|E|$'s.

A great deal of interest and high expectations were raised in this period from the introduction of the idea of applying the maximum-entropy method (MEM) to macromolecular crystallography. The MEM had enjoyed much success in the field of astronomy, where it had resulted in great improvements in the images of radio sources, and the processes of phase refinement and extension, where the aim is to improve a rather faulty image, seemed to lend themselves to the same treatment. Basically the method involves the maximization of an entropy function; the two which had been used are the Burg entropy function

$$S_B=\int\ln(\rho)\,\mathrm{d}V \tag{66}$$

and Jaynes' entropy function

$$S_J=-\int\rho\ln(\rho)\,\mathrm{d}V. \tag{67}$$

The idea of applying the MEM to crystallography originated with Narayan & Nityananda (1981, 1982) who favoured the Burg entropy function, but the most complete treatment was by Bricogne (1984) who used instead Jaynes's function. Bricogne demonstrated the application of the MEM to data from a small protein, but the data were calculated from the known structure and of a quality unattainable in practice. The gain in resolution was real and unmistakeable, but not impressive, and no application to an unknown structure with real data has been reported.

In fact Collins (1984) has shown that satisfying Sayre's equation is the equivalent of maximizing Jaynes's entropy function by changing only phases. Since, as is well known to crystallographers, phases, rather than magnitudes, contain most of the structural information it seems likely that a Sayre-equation-based method might confer most of the benefit of the MEM. Narayan & Nityananda (1982) have also shown that maximization of the Burg entropy is equivalent to satisfying the maximum-determinant condition of Tsoucaris (1970). It seems that entropy maximization is adding nothing completely new to the crystallographic scene and, since it involves a great deal of effort, perhaps nothing useful.

A range of work which is new, however, involves a combination of direct methods with either isomorphous-replacement data or data from anomalous scattering. For example, Hauptman (1982) has developed equations for estimating three-phase invariants if data are available from two isomorphous crystals. The method was applied to calculated error-free data for a protein (Hauptman, Potter & Weeks, 1982), giving reliable estimates for many thousands of three-phase invariants which were close to 0 or π. However, no

application to real data has been reported. If the heavy-atom structure is known then Fortier, Moore & Fraser (1985) have shown how to estimate three-phase invariants in the range 0 to 2π.

A simple rule for estimating three-phase invariants has been given by Karle (1984*a*, *b*) in the case that a heavy-atom derivative is available. If subscript p represents 'protein' and pH 'protein + heavy atom' then the rule is:

'If the sign of the product

$$(|F(\mathbf{h})_{pH}|-|F(\mathbf{h})_p|)(|F(\mathbf{k})_{pH}|-|F(\mathbf{k})_p|)$$
$$\times(|F(\mathbf{h}+\mathbf{k})_{pH}|-|F(\mathbf{h}+\mathbf{k})_p|)$$

is positive then the value of $\langle\varphi(\mathbf{h})-\varphi(\mathbf{k})-\varphi(\mathbf{h}+\mathbf{k})\rangle$ is close to zero while if the sign is negative the average invariant is close to π.' The average invariant in this case is the average of the eight values

$$\varphi(\mathbf{h})_A+\varphi(\mathbf{k})_B-\varphi(\mathbf{h}+\mathbf{k})_C,$$

where A, B and C can each be p or pH.

The sign will be a good indication for individual phase triplets if the corresponding value of

$$F(\mathbf{h})_A F(\mathbf{k})_B F(\mathbf{h}+\mathbf{k})_C$$

is large. If the heavy-atom structure is known then Karle (1986) has shown how it is possible to derive a good estimate between 0 and 2π for a three-phase invariant.

The single-isomorphous-replacement (SIR) method has the property that it leads to an ambiguity in estimating phases. Fan Hai-fu (1983) and Fan Hai-fu, Han Fu-son, Qian, Jin-zi & Yao Jia-xing (1984) have shown that by the use of probability formulae, based on the concepts of direct methods, the phase ambiguities could be broken. The same method could be applied to one-wavelength anomalous scattering (OAS) data which also give an ambiguity in phase. The method has been applied to real data for APP (avian pancreatic polypeptide) where the native protein and a derivative containing mercury were available. It was clear from the quality of the results obtained with OAS data of the Hg derivative that the structure would have been solved directly from this approach.

Other workers have developed techniques for allying direct methods with anomalous scattering data. Karle (1980, 1982*b*) has described a method whereby from data at several wavelengths linear equations can be set up and solved for quantities which include the magnitude squared of the structure factors of the anomalously scattering atoms alone. If these atoms form a simple structure then this can be solved from a Patterson map or by direct methods and thence the whole structure can be determined. The same general idea, where there is only one kind of anomalous scatterer and where only measurements at two wavelengths are needed, has been given by Cascarano, Giacovazzo, Peerdeman & Kroon (1982) and Woolfson (1984). Furthermore, Karle (1985) has shown that with only one type of anomalous scatterer the problem may be solved with data from one wavelength.

Finally, it must be mentioned that the first tentative steps are being taken to obtain phase information directly from experiment. When a crystal is positioned so that two diffracted beams are produced simultaneously then the profile of the beams can give information about the triple-sign product (21) or the three-phase invariant (29*a*). Post (1979) gave first results for centrosymmetric structures and Hümmer & Billy (1986) have given a very convincing demonstration of the technique applied to non-centrosymmetric crystals.

The crystal ball: 1987–

The purely physical methods to which reference has been made have so far only been applied to very simple structures, which could be easily solved by other methods, at a cost of experimental time which would be prohibitive for routine application. If the method can be made faster and applicable to more complicated structures then it will overtake all other procedures and make them redundant. Such progress cannot be expected in the short term but, given the history of development in crystallography, it must be seen as a possibility.

The hybrid methods, in which the concepts of direct methods are linked with SIR or OAS, look far more promising in the shorter term. There can be little doubt that such methods may begin to dominate in the next decade, especially with the increased availability of synchrotron sources so that the wavelengths used may be tuned to particular theoretical needs. The multiple-wavelength anomalous scattering method described by Karle could be very relevant here.

There is little doubt that the MEM is, from a theoretical point of view, the ultimate technique in terms of extracting information from the data. Bricogne (1984) has realized that, powerful although it is, it cannot be relied upon to lead from whatever starting point one has to the correct solution by a unique and unequivocal pathway. Thus, the algorithm he describes for *ab initio* structure solution is a multi-solution one with an expanding tree-like structure and where unlikely branches are lopped off from time to time to constrain the problem to manageable proportions. Nevertheless the process, which involves repeated applications of Fourier transformation, is a costly one and the present evidence does not suggest that the performance is commensurate with the cost.

As has been previously mentioned it seems that most of the theoretical benefit of the MEM is gained by finding a set of phases which satisfy the Sayre

equation – and that is exactly what is done by the application of the Sayre tangent formula. Whether or not the Sayre tangent formula can be a cheaper alternative to the MEM is yet to be explored; certainly it will be much cheaper to apply and may more than compensate for its slightly lesser theoretical power by enabling a much larger number of trials to be carried out in a multisolution mode.

The general practice in macromolecular work has been to devise single-path techniques – and this is understandable since the computational task for such structures can be quite massive. For small structures experience has shown that multisolution methods are far more powerful; indeed, it is not even certain that single-path methods will work at all for most structures. The only logical way to progress in a single-path method is to go forward in the most obvious direction at every stage, and this is rarely the correct path in the early stages. The same must certainly be true for macromolecular structures. With a high increase in available computer power and speed the right way to progress must be to incorporate a multisolution approach. As an example, one method of phase extension and refinement may be to accept estimates of low-resolution phases from isomorphous-replacement data and then populate the outer region of reciprocal space with random phases for selected large structure factors and refine with the Sayre tangent formula. In the last few cycles the inner phases can be allowed to change to fit in which the overall phase pattern. A number of trials will give phase sets reasonably consistent with the original isomorphous-replacement data and satisfying Sayre's equation. Whether or not this idea will work or be as good as might be achieved, with more effort, by the MEM remains to be seen. It does have some of the character of Yao's method of building on a fragment, which is also a multisolution approach to using partial information to derive the whole structure. In one case the information which grows is the number of atomic positions; in the second case it is the resolution.

A multisolution approach may enable *ab initio* solutions of proteins to be contemplated. At the lowest resolution when an electron density map begins to show recognizably correct features, the number of data being deployed is not too large – comparable indeed to the number used to solve a normal small structure. If suitable phase relationships can be found, based on whatever physical criteria are appropriate at the resolution in question, then it is conceivable that a finite number of low-resolution maps could be generated, including a correct one. Phase refinement and extension might then take over.

A famous nuclear physicist was once asked in the early 1950's what he thought nuclear physicists would be doing ten years thereafter. His response was that if he knew then he would be doing it now. If the author knew what crystallographers would be doing at the end of the century you may be sure that *he* would be doing it now!

Thanks are due to referees whose comments led to considerable improvement in the article.

References

BRICOGNE, G. (1984). *Acta Cryst.* **A40**, 410-445.

CASCARANO, G., GIACOVAZZO, C., PEERDEMAN, A. F. & KROON, J. (1982). *Acta Cryst.* **A38**, 710-717.

COCHRAN, W. (1952). *Acta Cryst.* **5**, 65-67.

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

COCHRAN, W. & DOUGLAS, A. S. (1955). *Proc. R. Soc. London Ser. A*, **227**, 486-500.

COCHRAN, W. & PENFOLD, B. R. (1952). *Acta Cryst.* **5**, 644-654.

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

COLLINS, D. M. (1984). *Acta Cryst.* **A40**, C426.

CUTFIELD, J. F., DODSON, E. J., DODSON, G. G., HODGKIN, D. C., ISAACS, N. W., SAKABE, K. & SAKABE, N. (1975). *Acta Cryst.* **A31**, S21.

DEBAERDEMAEKER, T., TATE, C. & WOOLFSON, M. M. (1985). *Acta Cryst.* **A41**, 286-290.

DEBAERDEMAEKER, T. & WOOLFSON, M. M. (1983). *Acta Cryst.* **A39**, 193-196.

DECLERCQ, J.-P., GERMAIN, G., VAN MEERSSCHE, M., HULL, S. E. & IRWIN, M. J. (1978). *Acta Cryst.* **B34**, 3644-3648.

DECLERCQ, J.-P., GERMAIN, G. & WOOLFSON, M. M. (1975). *Acta Cryst.* **A31**, 367-372.

DE TITTA, G. T., EDMONDS, J. W., LANGS, D. A. & HAUPTMAN, H. (1975). *Acta Cryst.* **A31**, 472-479.

FAN HAI-FU (1983). *Methods and Applications in Crystallographic Computing*, edited by S. R. HALL & T. ASHIDA, p. 482. Oxford: Clarendon Press.

FAN HAI-FU, HAN FU-SON, QIAN JIN-ZI & YAO JIA-XING (1984). *Acta Cryst.* **A40**, 489-495.

FAN HAI-FU, YAO JIA-XING, MAIN, P. & WOOLFSON, M. M. (1983). *Acta Cryst.* **A39**, 566-569.

FORTIER, S., MOORE, N. J. & FRASER, M. E. (1985). *Acta Cryst.* **A41**, 571-577.

GERMAIN, G., MAIN, P. & WOOLFSON, M. M. (1970). *Acta Cryst.* **B26**, 274-285.

GERMAIN, G. & WOOLFSON, M. M. (1968). *Acta Cryst.* **B24**, 91-96.

GIACOVAZZO, C. (1977). *Acta Cryst.* **A33**, 933-944.

GILLIS, J. (1948). *Acta Cryst.* **1**, 76-80.

GILMORE, C. J. (1984). *J. Appl. Cryst.* **17**, 42-46.

GOEDKOOP, J. A. (1950). *Acta Cryst.* **3**, 374-378.

GRANT, D. F., HOWELLS, R. G. & ROGERS, D. (1957). *Acta Cryst.* **10**, 489-497.

HARKER, D. & KASPER, J. S. (1948). *Acta Cryst.* **1**, 70-75.

HAUPTMAN, H. (1975). *Acta Cryst.* **A31**, 671-679.

HAUPTMAN, H. (1982). *Acta Cryst.* **A38**, 289-294.

HAUPTMAN, H., FISHER, J., HANCOCK, H. & NORTON, D. (1969). *Acta Cryst.* **B25**, 811-814.

HAUPTMAN, H. & KARLE, J. (1950*a*). *Phys. Rev.* **77**, 491-499.

HAUPTMAN, H. & KARLE, J. (1950*b*). *Acta Cryst.* **3**, 478.

HAUPTMAN, H. & KARLE, J. (1953). *The Solution of the Phase Problem*: I. *The Centrosymmetric Crystal. Am. Crystallogr. Assoc. Monogr.* No. 3. Wilmington: The Letter Shop.

HAUPTMAN, H. & KARLE, J. (1956). *Acta Cryst.* **9**, 45-55.

HAUPTMAN, H. & KARLE, J. (1959). *Acta Cryst.* **12**, 93-97.

HAUPTMAN, H., POTTER, S. & WEEKS, C. M. (1982). *Acta Cryst.* **A38**, 294-300.

HULL, S. E. & IRWIN, M. J. (1978). *Acta Cryst.* **A34**, 863-870.

HÜMMER, K. & BILLY, H. (1986). *Acta Cryst.* **A42**, 127-133.

KARLE, I. L., HAUPTMAN, H., KARLE, J. & WING, A. B. (1958). *Acta Cryst.* **11**, 257-263.

KARLE, I. L. & KARLE, J. (1963). *Acta Cryst.* **16**, 969-975.

KARLE, I. L. & KARLE, J. (1964*a*). *Acta Cryst.* **17**, 835-841.

KARLE, I. L. & KARLE, J. (1964*b*). *Acta Cryst.* **17**, 1356-1360.
KARLE, J. (1968). *Acta Cryst.* **B24**, 182-186.
KARLE, J. (1971). *Acta Cryst.* **B27**, 2063-2065.
KARLE, J. (1980). *Int. J. Quantum Chem.* **7**, 357-367.
KARLE, J. (1982*a*). *Acta Cryst.* **A38**, 327-333.
KARLE, J. (1982*b*). *Computational Crystallography*, edited by D. SAYRE, pp. 174-200. Oxford: Clarendon Press.
KARLE, J. (1984*a*). *Acta Cryst.* **A40**, 526-531.
KARLE, J. (1984*b*). *Methods and Applications in Crystallographic Computing*, edited by S. R. HALL & T. ASHIDA, pp. 120-140. Oxford: Clarendon Press.
KARLE, J. (1985). *Acta Cryst.* **A41**, 387-394.
KARLE, J. (1986). *Acta Cryst.* **A42**, 246-253.
KARLE, J. & HAUPTMAN, H. (1950). *Acta Cryst.* **3**, 181-187.
KARLE, J. & HAUPTMAN, H. (1956). *Acta Cryst.* **9**, 635-651.
KARLE, J. & HAUPTMAN, H. (1957). *Acta Cryst.* **10**, 515-524.
KARLE, J. & HAUPTMAN, H. (1961). *Acta Cryst.* **14**, 217-223.
KLUG, A. (1958). *Acta Cryst.* **11**, 515-543.
MAIN, P. (1977). *Acta Cryst.* **A33**, 750-757.
MAIN, P. & HULL, S. E. (1978). *Acta Cryst.* **A34**, 353-361.
MAUGUEN, Y. (1979). Thèse. Univ. de Paris VI.
NARAYAN, R. & NITYANANDA, R. (1981). *Curr. Sci.* **50**, 168-170.
NARAYAN, R. & NITYANANDA, R. (1982). *Acta Cryst.* **A38**, 122-128.
OKAYA, Y. & NITTA, I. (1952). *Acta Cryst.* **5**, 564-570, 687-688.
OVERBEEK, A. R. & SCHENK, H. (1978). In *Computing in Crystallography*, edited by H. SCHENK, R. OLTHOF-HAZEKAMP, H. VAN KONINGSVELD & G. C. BASSI. Delft Univ. Press.
PATTERSON, A. L. (1934). *Phys. Rev.* **46**, 372-376.
POST, B. (1979). *Acta Cryst.* **A35**, 17-21.
RANGO, C. DE, TSOUCARIS, G. & ZELWER, C. (1974). *Acta Cryst.* **A30**, 342-353.
SAKURAI, K. (1952). *Acta Cryst.* **5**, 546-548, 697.
SAYRE, D. (1952). *Acta Cryst.* **5**, 60-65.
SAYRE, D. (1972). *Acta Cryst.* **A28**, 210-212.
SAYRE, D. (1974). *Acta Cryst.* **A30**, 180-184.
SCHENK, H. (1973*a*). *Acta Cryst.* **A29**, 77-82.
SCHENK, H. (1973*b*). *Acta Cryst.* **A29**, 480-481.
SCHENK, H. & DE JONG, J. G. H. (1973). *Acta Cryst.* **A29**, 31-34.
SHELDRICK, G. M. (1975). *SHELX*. Program for crystal structure determination. Univ. of Cambridge, England.
TOEPLITZ, O. (1911). *Rend. Circ. Mat. Palermo*, **32**, 191-192.
TSOUCARIS, G. (1970). *Acta Cryst.* **A26**, 492-499.
WHITE, P. & WOOLFSON, M. M. (1975). *Acta Cryst.* **A31**, 53-56.
WOOLFSON, M. M. (1954). *Acta Cryst.* **7**, 61-67.
WOOLFSON, M. M. (1958). *Acta Cryst.* **11**, 4-6.
WOOLFSON, M. M. (1961). *Direct Methods in Crystallography*. Oxford Univ. Press.
WOOLFSON, M. M. (1977). *Acta Cryst.* **A33**, 219-225.
WOOLFSON, M. M. (1984). *Acta Cryst.* **A40**, 32-34.
WRINCH, D. M. (1939). *Philos. Mag.* **27**, 98.
YAO JIA-XING (1981). *Acta Cryst.* **A37**, 642-644.
YAO JIA-XING (1983). *Acta Cryst.* **A39**, 35-37.
ZACHARIASEN, W. H. (1952). *Acta Cryst.* **5**, 68-73.

Acta Cryst. (1987). **A43**, 612-616

Multi-symmetric Close Packings of Equal Spheres on the Spherical Surface

BY T. TARNAI

Hungarian Institute for Building Science, Budapest, Dávid F.u. 6, *H*-1113 *Hungary*

AND ZS. GÁSPÁR

Department of Civil Engineering Mechanics, Technical University of Budapest, Budapest, Müegyetem rkp. 3, *H*-1111 *Hungary*

(*Received* 26 *May* 1986; *accepted* 22 *January* 1987)

Abstract

An analysis is presented for the Tammes problem: how must n points be distributed on the surface of a sphere in order that the minimum angular distance between any two of the points be a maximum? With the analogy of the capsid structure of small 'spherical' viruses, locally extremal arrangements are constructed in tetrahedral, octahedral and icosahedral symmetry. Thirty arrangements defined by four packing sequences are investigated. By the applied construction process, novel locally extremal configurations for $n = 78$, 96, 108, 144, 150, 192, 198, 270, 360, 372, 480, 492 and improvable configurations for $n = 114$, 282 are obtained. A table is given of the investigated arrangements; most of them are putative solutions of the Tammes problem.

Introduction

Consider the problem of the closest packing of n equal non-intersecting spheres on the spherical surface investigated by Mackay, Finney & Gotoh (1977).

This 'hard-sphere' problem, mentioned as the Fejes problem (Fejes Tóth, 1972) by Mackay, Finney & Gotoh (1977) but better known as the Tammes problem (Tammes, 1930; Fejes Tóth, 1964), has several equivalent formulations. Melnyk, Knop & Smith (1977) enumerated the different formulations of this purely geometrical problem but presented also a physical interpretation of it as an extreme case of finding equilibrium configuration where n points on the surface of a sphere repel each other according to the inverse power law. Namely, when the exponent of the power tends to infinity, the smallest distance

0108-7673/87/050612-05$01.50