Crystallographic Computing 5

FROM CHEMISTRY TO BIOLOGY

Papers presented at the
International School on Crystallographic Computing
held at
Bischenberg, France
29 July – 5 August 1990

Edited by

D. Moras
A. D. Podjarny
and
J. C. Thierry

Laboratoire de Cristallographie Biologique
Institut de Biologie Moleculaire et Cellulaire du CNRS
Strasbourg, France

INTERNATIONAL UNION OF CRYSTALLOGRAPHY
OXFORD UNIVERSITY PRESS
1991

24

A minimal principle in the phase problem

H.A. Hauptman

1. Introduction

1.1. The Phase Problem

The intensities of a sufficient number of x-ray diffraction maxima determine a crystal structure. The available intensities usually exceed the number of parameters needed to describe the structure. From these intensities a set of numbers $|E_{\mathbf{H}}|$ can be derived, one corresponding to each intensity. However the elucidation of the crystal structure requires also a knowledge of the complex numbers $E_{\mathbf{H}} = |E_{\mathbf{H}}|\exp(i\phi_{\mathbf{H}})$, the normalized structure factors, of which only the magnitudes $|E_{\mathbf{H}}|$ can be determined from experiment. Thus a "phase" $\phi_{\mathbf{H}}$, unobtainable from the diffraction experiment, must be assigned to each $|E_{\mathbf{H}}|$, and the problem of determining the phases when only the magnitudes $|E_{\mathbf{H}}|$ are known is called "the phase problem". Owing to the known atomicity of crystal structures and the redundancy of observed magnitudes $|E_{\mathbf{H}}|$, the phase problem is solvable in principle.

Direct methods are those procedures which attempt to solve the phase problem by reconstructing the lost phase information directly from the observed structure factor amplitudes. These methods rely on the existence of relationships among the structure factors which express the values of certain linear combinations of phases, called structure invariants, in terms of normalized structure factor magnitudes.

1.2. The Normalized Structure Factors E

In the equal atom case the normalized structure factors are defined by

$$E_H = |E_H| \exp(i\phi_H) = \frac{1}{N^{1/2}} \sum_{j=1}^{N} \exp\left(2\pi i \mathbf{H}\cdot\mathbf{r}_j\right) \qquad (1)$$

where H is an arbitrary reciprocal lattice vector, N is the number of atoms in the unit cell and $\mathbf{r}_j$ is the position vector of the atom labeled j. The magnitudes $|E|$ are directly obtainable from the diffraction intensities, but the phases ϕ are lost in the diffraction experiment.

1.3. The Structure Invariants

Although the values of the individual phases are known to depend on the structure and the choice of origin, there exist certain linear combinations of the phases whose values are determined by the structure alone and are independent of the choice of origin. These linear combinations of the phases are called the structure invariants. The most important classes of structure invariants, and the only ones to be used here, are the three-phase structure invariants (triplets),

$$\phi_{HK} = \phi_H + \phi_K + \phi_{-H-K} \,, \qquad (2)$$

and the four-phase structure invariants (quartets),

$$\phi_{LMN} = \phi_L + \phi_M + \phi_N + \phi_{-L-M-N} \,. \qquad (3)$$

2. The Probabilistic Background

It is assumed that the position vectors $\mathbf{r}_j$ are random variables which are uniformly and independently distributed. Then the structure invariants, as functions of random variables via Eqs. (1)-(3), are themselves random variables, and their conditional probability

distributions, assuming as known certain magnitudes $|E|$, may then be found.

2.1. The Conditional Probability Distribution of the Triplet

For fixed reciprocal lattice vectors H and K, the conditional probability distribution of the triplet ϕ_{HK} [Eq. (2)], assuming as known the three magnitudes

$$|E_H|, \ |E_K|, \ |E_{H+K}|, \tag{4}$$

is known to be

$$P\left(\phi|A_{HK}\right) = \frac{1}{2\pi I_0(A_{HK})} \exp\left(A_{HK} \cos \phi\right) \tag{5}$$

where

$$A_{HK} = \frac{2}{N^{1/2}} |E_H E_K E_{H+K}|, \tag{6}$$

and I is the Modified Bessel Function. Eq. (5) implies that the mode of ϕ_{HK} is zero, the conditional expectation value of cosine ϕ_{HK} given A_{HK}, is

$$\varepsilon\left(\cos \phi_{HK}|A_{HK}\right) = \frac{I_1(A_{HK})}{I_0(A_{HK})} > 0, \tag{7}$$

and that the larger the value of A_{HK} the smaller is the conditional variance of cos ϕ_{HK}, given A_{HK}. It is to be stressed that the conditional expected value of the cosine, Eq. (7), is always positive since $A_{HK}>0$.

2.2. The Quartet

For fixed reciprocal lattice vectors L, M, and N, the conditional probability distribution of the quartet ϕ_{LMN} [Eq. (3)], assuming as

known the seven magnitudes

$$|E_L|,\ |E_M|,\ |E_N|,\ |E_{L+M+N}|,\ |E_{L+M}|,\ |E_{M+N}|,\ |E_{N+L}|, \tag{8}$$

is now known. For our purpose it will be sufficient to use the approximation

$$P\left(\phi|B_{LMN}\right) = \frac{1}{2\pi I_0(B_{LMN})} \exp\left(B_{LMN} \cos \phi\right) \tag{9}$$

where

$$B_{LMN} = \frac{2}{N}|E_L E_M E_N E_{L+M+N}| \left\{ |E_{L+M}|^2 + |E_{M+N}|^2 + |E_{N+L}|^2 - 2 \right\}. \tag{10}$$

As in Eq. (8) we now find

$$\varepsilon\left(\cos \phi_{LMN}|B_{LMN}\right) = \frac{I_1(B_{LMN})}{I_0(B_{LMN})} \tag{11}$$

and the larger the value of $|B_{LMN}|$ the smaller the conditional variance of $\cos \phi_{LMN}$, given B_{LMN}. In sharp contrast to Eq. (7), the conditional expected value of the cosine [Eq. (11)] is now positive or negative according as

$$B_{LMN} \gtrless 0, \tag{12}$$

i.e., in view of (10), according as the "cross-terms" $|E_{L+M}|$, $|E_{M+N}|$, and $|E_{N+L}|$ are mostly large or mostly small, respectively. Those quartets for which $B_{LMN}<0$ are known as negative quartets because their cosines are probably negative. The special importance of the negative quartets will be emphasized in the sequel. It is to be stressed that it is only the expected values of the cosines of the negative quartets which are negative; not all cosines of negative quartets are necessarily negative.

3. The Minimal Principle

3.1. The Heuristic Background

It is assumed that a crystal structure S in the space group G and consisting of N identical atoms in the asymmetric unit is fixed, but unknown, that the magnitudes $|E|$ of the normalized structure factors E are known, and that a sufficiently large base of phases, corresponding to the largest magnitudes $|E|$, is specified.

The mode of the triplet distribution [Eq. (5)] is zero and the variance of the cosine is small if A_{HK} (Eq. 6) is large. In this way one obtains the estimate for the triplet ϕ_{HK} [Eq. (6)]:

$$\phi_{HK} = \phi_H + \phi_K + \phi_{-H-K} \tilde{=} 0 \qquad (13)$$

which is particularly good in the favorable case that A_{HK}, [Eq. (6)], is large, i.e. that $|E_H|$, $|E_K|$, and $|E_{H+K}|$ are all large. The estimate given by Eq. (13) is one of the cornerstones of current techniques of direct methods. It is surprising how useful Eq. (13) has proven to be in the applications especially since it yields only the zero estimate of the triplet, and only those estimates are reliable for which $|E_H|$, $|E_K|$, and $|E_{H+K}|$ are all large. Clearly the coefficient $2/N^{1/2}$ in Eq. (6), and therefore A_{HK} as well, both decrease with increasing N, i.e. with increasing structural complexity. Hence the relationship [Eq. (13)] becomes increasingly unreliable for larger structures, and the traditional step-by-step sequential direct methods procedures based on Eq. (13) eventually fail.

By their almost exclusive reliance on the triplet relationship [Eq. (13)], the traditional direct methods techniques do not fully exploit our detailed knowledge of the triplet distribution [Eq. (5)] and ignore almost completely the quartet distribution [Eq. (9)]. It is now

proposed to determine the values of the phases ϕ in such a way that they generate triplets and quartets which, for each fixed value of A_{HK} or B_{LMN}, have distributions which agree with their theoretical distributions, Eqs. (5) or (9), respectively. More specifically, one determines the values of a set of phases as those which generate triplets ϕ_{HK} and quartets ϕ_{LMN} whose cosines have, for each fixed value of A_{HK} and B_{LMN}, conditional expectation values and variances in agreement with their known theoretical values. In this way one exploits more effectively our knowledge of the triplet and quartet distributions. In this connection it should be noted that, for a sufficiently large basis set of phases, say more than some 300 phases in the base, the number of structure invariants which they generate exceeds by far (two or three orders of magnitude at least) the number of unknown phases ϕ. Owing to this great redundancy, a large number of identities among the structure invariants, equal to the difference between the number of structure invariants and the number of phases, must be satisfied. An important aspect of our present formulation is that all identities among the structure invariants, which must of necessity hold, will in fact be satisfied.

3.2. Triplets

In view of Eq. (7) and the previous discussion one now replaces the zero estimate [Eq. (13)] of the triplet ϕ_{HK} [Eq. (2)] by the estimate

$$\cos \phi_{HK} \simeq \frac{I_1(A_{HK})}{I_0(A_{HK})} \tag{14}$$

and expects that the smaller the variance, that is the larger A_{HK}, the more reliable the estimate [Eq. (14)] will be. Hence one is led to construct the function, determined by the known magnitudes $|E|$,

$$R = \frac{1}{\sum_{H,K} A_{HK}} \sum_{H,K} A_{HK} \left[\cos \phi_{HK} - \frac{I_1(A_{HK})}{I_0(A_{HK})} \right]^2 \qquad (15)$$

which is seen to be a function of all those triplets ϕ_{HK} which are generated by a prescribed set of phases $\{\phi\}$. Recall that if the basis set of phases $\{\phi\}$ is sufficiently large then there are many more structure invariants ϕ_{HK} than individual phases ϕ, and a myriad of identities among these structure invariants must, of necessity, then be satisfied. It is therefore natural to suppose that that set of values for the structure invariants ϕ_{HK} is best which minimizes the residual R, [Eq. (15)], subject to the constraint that all identities among the structure invariants are in fact satisfied.

Since the triplets ϕ_{HK} are defined by Eq. (2) as functions of the individual phases ϕ, Eq. (15) defines R implicitly as a function of the individual phases. One therefore naturally anticipates that that set of values for the individual phases is best which minimizes the residual R, Eq. (15), nov regarded as a function of the individual phases ϕ. The advantage of this formulation is that all identities among the structure invariants will then automatically be satisfied, and it is unnecessary to define in further detail what the nature of these identities must be.

3.3. The Minimal Principle for Triplets

In order to derive the conditions under which the formulation of the minimal principle given in the previous paragraph is valid, one first defines R_T as the value of R [Eq. (15)] obtained when the phases are equal to their true values for some choice of origin and enantiomorph. One then defines R_R as the value of R when the phases are assigned values at random so that

$$\left\langle \cos \phi_{HK} \right\rangle_{H,K} = \left\langle \cos 2\phi_{HK} \right\rangle_{H,K} = 0. \tag{16}$$

With these definitons for R_T and R_R it may then be shown that

$$R_T < \frac{1}{2} < R_R. \tag{17}$$

In spite of the inequalities, [Eq. (17)], it still does not follow that R_T is the global minimum of R because, as it turns out, there exist certain special values of the phases which yield values for R [Eq. (15)] even less than R_T. It is for this reason that one must introduce the quartets, in particular the negative quartets, in order to insure that the global minimum yields the true values for the phases.

3.4. The Minimal Principle

In view of the previous paragraph one incorporates the negative quartets into the definition of the minimal function as follows:

$$R = \frac{\sum_{H,K} A_{HK}\left(\cos \phi_{HK} - \frac{I_1(A_{HK})}{I_0(A_{HK})}\right)^2 + \sum_{L,M,N} |B_{LMN}| \left(\cos \phi_{LMN} - \frac{I_1(B_{LMN})}{I_0(B_{LMN})}\right)^2}{\sum_{H,K} A_{HK} + \sum_{L,M,N} |B_{LMN}|} \tag{18}$$

where the double sum is taken over all triplets ϕ_{HK} generated by the basis set of phases, and the triple sum is taken over all the negative quartets ϕ_{LMN}, that is those for which $B_{LMN}<0$.

There remains one final word of caution. In general the number of phases in the base must exceed the number of parameters, 3N, needed to define the crystal structure. Hence the phases themselves are not independent variables but must themselves satisfy a number of identities equal to the number of phases diminished by 3N. Hence the final

formulation of the minimal principle is simply that those phases are correct which minimize R [Eq. (18)] subject to the constraint that all identities among the phases be satisfied.

Since, for fixed origin, the crystal structure determines the values of all the phases, the function R may be regarded as a function of structures, T, and the minimal principle asserts that the structure T which minimizes R is the desired structure S.

Strategies for finding the global minimum of R which employ a modified simulated annealing algorithm have been devised. The method has been used (employing also a small number of positive quartets) to solve the two (previously known) structures in the space group $P2_12_12_1$: $C_{18}H_{28}O_{11}$, Z=4, 2138 reflections, 300 phases in the base, number of triplets = 7110, number of negative quartets = 139,290, number of positive quartets = 8805, error free data; and $C_{60}N_6O_{18}H_{102}$, Z=4, 5024 reflections, 400 phases in the base, number of triplets = 6514, number of negative quartets = 90,001, number of positive quartets = 2880, experimental data.

In this work the only structure invariants used were the triplets and the (mostly negative) quartets the probabilistic theories of which are well known. It is not yet known whether the method described here will be useful for structures of much greater complexity because it is still not clear how rapidly the size of the phase base must increase as a function of increasing N.