The Phase Problem: A Problem in Constrained Global Optimization

Herbert A. Hauptman
Hauptman-Woodward Medical Research Institute, Inc.
73 High Street, Buffalo, New York 14203-1196
hauptman@hwi.buffalo.edu

Abstract

It is now almost 200 years since Gauss, a teenager at the time, formulated his famous principle of least-squares and used it to determine, for the first time, the orbit of one of the asteroids, a problem which had defeated astronomers for years. When applied to the crystallographic phase problem, least-squares leads directly to the formulation of the minimal principle, which effectively replaces the phase problem by one of constrained global minimization. Shake-and-Bake, the computer software package which implements this formulation of the phase problem, provides a completely automatic solution of this problem. Shake-and-Bake requires that diffraction intensities to a resolution of 1.2Å, at least, be available. Structures having as many as 600 independent non-hydrogen atoms have been routinely solved in this way; the ultimate potential of the method is still not known.

When single-wavelength anomalous scattering (SAS) diffraction data are available, the phase problem may again be formulated as a problem in global optimization. Although the objective function has a myriad of local maxima, its global maxima, never more than two, are readily accessible and easily identified by virtue of their isolation. The ability to determine the global maxima of the objective function represents the latest and most successful attempt to go directly from the known probabilistic estimates of the three-phase structure invariants to the values of the individual phases. The relationship between the maxima of the objective function and the solutions of the newly formulated system of SAS tangent equations plays a key role in this development,

1 Introduction

The techniques of modern probability theory lead to the joint probability distributions of arbitrary collections of diffraction intensities and their corresponding phases. These distributions constitute the foundation on which direct methods are based. They have provided the unifying thread from the beginning, *ca* 1950, until the present time. They have led, in particular to the (first) minimal principle [1-3] which has found expression in the ***Shake-and-Bake*** formalism [4,5], a computer program which provides a completely automatic solution to the phase problem, *ab initio*, provided that diffraction data to at least 1.2Å are available. Our experience shows that structures having as many as 600 independent non-hydrogen atoms are routinely accessible to this approach and suggests that its ultimate potential is greater still.

One naturally anticipates that, with the availability of single-wavelength anomalous scattering data, the ability to determine phases *ab initio* will be strengthened. This expectation is in fact realized here. Specifically, an SAS maximal principle is formulated which, even though unconstrained, nevertheless strengthens the earlier minimal principle by incorporating SAS estimates of the cosines and sines of the three-phase structure invariants. The initial applications show that, in this way, the phase problem is solvable, *ab initio*, even for macromolecules, when SAS diffraction data alone are available at a resolution of about 2.5Å.

This work represents the latest attempt to go directly from known estimates of the three-phase structure invariants to the values of the individual phases (*cf.*, for example, [7,8]). However, instead of attempting to solve by least-squares a redundant system of linear equations, as was done in the earlier work, the formulation presented here transforms the problem into one of global optimization, a problem with a surprisingly easy solution. Furthermore, as shown by the initial applications, briefly described here, these results represent a substantial improvement over the earlier work.

2 The Non-SAS Case

2.1 The Nature of the Constraints

For a structure consisting of N identical atoms in the unit cell the normalized structure factor E_H is defined by

$$E_H = \frac{1}{N^{1/2}} \sum_{j=1}^{N} \exp\left(2\pi i H \cdot r_j\right). \qquad (2.1)$$

Since the number of equations (2.1) exceeds by far the number of unknown atomic position vectors r_j, elimination of the r_j's leads to a system of equations among the normalized structure factors E_H alone:

$$F(E) = 0, \qquad (2.2)$$

each function F of which may be written as a function of known magnitudes |E| and unknown phases ϕ:

$$F(E) \rightarrow F(|E|;\phi) = 0. \qquad (2.3)$$

The system of equations (2.3) leads directly to a system of identities among the phases:

$$F(|E|;\phi) \rightarrow G(\phi/|E|) = 0 \qquad (2.4)$$

where the G's are functions of the phases ϕ that depend upon the known parameters |E|. One infers then that the phases are, of necessity, constrained to satisfy the system of identities (2.4).

2.2 The Probabilistic Background

If one assumes that the atomic position vectors r_j are the primitive random variables, uniformly and independently distributed in the unit cell, then the normalized structure factors E, as functions of the primitive random variable r_j, Eq. (2.1), are themselves random variables. Since the magnitudes |E| are obtainable from the diffraction experiment, the phases ϕ may be regarded as random variables, the conditional probability distribution of any collection of which, assuming as given appropriate magnitudes |E|, may be found by standard techniques. Thus the conditional probability distribution of the triplet

$$\phi_{\mathbf{HK}} = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}}, \qquad (2.5)$$

given the three magnitudes

$$|E_{\mathbf{H}}|,\ |E_{\mathbf{K}}|,\ |E_{\mathbf{H}+\mathbf{K}}|, \qquad (2.6)$$

is known to be

$$P(\Phi/\kappa_{\mathbf{HK}}) = \frac{1}{2\pi I_0(\kappa_{\mathbf{HK}})} \exp(\kappa_{\mathbf{HK}} \cos\Phi_{\mathbf{HK}}), \qquad (2.7)$$

where Φ represents the triplet $\phi_{\mathbf{HK}}$ (Eq. (2.5)), $\kappa_{\mathbf{HK}}$ is given by

$$\kappa_{\mathbf{HK}} = \frac{2}{N^{1/2}} |E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{H}+\mathbf{K}}|, \qquad (2.8)$$

and I_0 is the Modified Bessel Function. From Eq. (2.7) one finds the expected value of cos $\phi_{\mathbf{HK}}$ to be

$$\mathcal{E}(\cos\phi_{\mathbf{HK}}) = \frac{I_1(\kappa_{\mathbf{HK}})}{I_0(\kappa_{\mathbf{HK}})}, \qquad (2.9)$$

where I_1 is the Modified Bessel Function, and infers that the larger the value of the parameter $\kappa_{\mathbf{HK}}$ the smaller is the variance of the cosine.

2.3 The Minimal Principle

In view of § 2.2 one defines the minimal function, a function of the phases ϕ, by means of

$$m(\phi) = \frac{\sum_{\mathbf{H,K}} \kappa_{\mathbf{HK}} \left\{ \cos\phi_{\mathbf{HK}} - \frac{I_1(\kappa_{\mathbf{HK}})}{I_0(\kappa_{\mathbf{HK}})} \right\}^2}{\sum_{\mathbf{H,K}} \kappa_{\mathbf{HK}}} \qquad (2.10)$$

and infers that the global minimum of m(ϕ), where the phases ϕ are constrained to satisfy the system of identities (2.4), yields the true values of the phases for some choice of origin and enantiomorph (the minimal principle). In this way the phase problem is formulated as one of *constrained* global minimization, with emphasis on the word *constrained*: the *unconstrained* global minimum of m(ϕ) does not solve the phase problem; the *constrained* global minimum does. The reader is referred to DeTitta *et al.* [3], Weeks *et al.* [4], and Miller *et al.* [5] for further details, in particular how the computer program Shake-and-Bake converges to the constrained global minimum of m(ϕ).

3 The SAS Case

Once again the probabilistic theory of the (three-phase) structure invariants, initiated in the SAS case in 1982 [6], plays the central role. It should perhaps be stressed at the outset that, owing now to the breakdown of Friedel's Law and contrary to all earlier belief, unique values for all the structure invariants in the whole interval from 0 to 2π are determined since the enantiomorph is fixed by the observed magnitudes |E|. It is believed that the ability to fix the enantiomorph, *ab initio*, accounts for the unexpected result described here (§3.5).

The approach adopted here is similar to that used in the derivation of the minimal principle but is suitably modified in order to take into account the availability of the SAS diffraction data. Not only is one led in this way to the SAS maximal principle, but an important connection with the SAS tangent formula, the analogue of the traditional tangent formula, is established. Two remarkable properties of the SAS maximal function emerge: (a) the easy accessibility and ready identification of its global maxima and (b) the isolated character of these maxima.

3.1 The Probabilistic Background

With the assumption that SAS diffraction data are available, the conditional probability distribution P(ϕ) of the triplet

$$\phi_{\mathbf{HK}} = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}}, \qquad (3.1)$$

given the six magnitudes

$$|E_{\mathbf{H}}|,\ |E_{-\mathbf{H}}|,\ |E_{\mathbf{K}}|,\ |E_{-\mathbf{K}}|,\ |E_{\mathbf{H}+\mathbf{K}}|,\ |E_{-\mathbf{H}-\mathbf{K}}|, \qquad (3.2)$$

is known to be [6],

$$P(\phi) = \left[2\pi I_0(A_{\mathbf{HK}})\right]^{-1} \exp\left\{A_{\mathbf{HK}} \cos(\phi - \omega_{\mathbf{HK}})\right\} \qquad (3.3)$$

in which I_0 is the Modified Bessel Function and $A_{\mathbf{HK}}$ and $\omega_{\mathbf{HK}}$ are expressed in terms of the six magnitudes (3.2) and the (presumed known) complex-valued atomic scattering factors f̃. Hence $A_{\mathbf{HK}}(>0)$ and $\omega_{\mathbf{HK}}$ are here assumed to be known for every pair (**H,K**). Note that, owing to the breakdown of Friedel's Law, the six magnitudes (3.2) are, in general, distinct.

In view of Eq. (3.3), the most probable value of $\phi_{\mathbf{HK}}$ is $\omega_{\mathbf{HK}}$, and the larger the value of $A_{\mathbf{HK}}$ the better is this estimate of $\phi_{\mathbf{HK}}$:

$$\phi_{\mathbf{HK}} = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}} \approx \omega_{\mathbf{HK}}. \qquad (3.4)$$

3.2 The SAS Maximal Principle

In exact analogy with the derivation of the minimal principle in the non-SAS case, one now defines the SAS maximal function $M(\phi)$ by means of

$$M(\phi) = \frac{\sum_{\mathbf{H,K}} A_{\mathbf{HK}} \cos(\phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}} - \omega_{\mathbf{HK}})}{\sum_{\mathbf{H,K}} A_{\mathbf{HK}}} \qquad (3.5)$$

and infers that the global maximum of $M(\phi)$ yields the true values of the phases for some choice of origin (the SAS maximal principle). There remains the problem of finding the global maximum of $M(\phi)$, a problem presumed to be difficult by virtue of the existence of a myriad of local maxima of $M(\phi)$. The solution, however, turns out to be unexpectedly straightforward. How the problem is solved via the system of SAS tangent equations is described next.

3.3 The System of SAS Tangent Equations

Eq. (3.4) implies the SAS tangent formula: For each fixed value of the reciprocal lattice vector H

$$\tan\phi_{\mathbf{H}} = \frac{\sum_{\mathbf{K}} A_{\mathbf{HK}} \sin(\omega_{\mathbf{HK}} - \phi_{\mathbf{K}} - \phi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} A_{\mathbf{HK}} \cos(\omega_{\mathbf{HK}} - \phi_{\mathbf{K}} - \phi_{-\mathbf{H}-\mathbf{K}})} \qquad (3.6)$$

where the sign of $\sin\phi_{\mathbf{H}}$ is given by the numerator of (3.6) and the sign of $\cos\phi_{\mathbf{H}}$ by the denominator. Thus the tangent formula determines a unique value for $\phi_{\mathbf{H}}$ when the values of all other phases are assumed to be known.

Although (3.6) differs from the standard tangent formula only in the presence of the non-zero estimates $\omega_{\mathbf{HK}}$ of the three-phase structure invariants $\phi_{\mathbf{HK}}$, as well as the completely different set of weights $A_{\mathbf{HK}}$, these differences are of fundamental importance in the applications. As described in the sequel, the present formulation solves the phase problem for macromolecules in the SAS case with diffraction data to 2.5Å resolution; not surprisingly, the same claim cannot be made for the standard tangent formula when used in the same way.

3.4 The Maximal Property of the SAS Tangent Formula

Fix the reciprocal lattice vector H. Assume that the values of all phases other than $\phi_{\mathbf{H}}$ are specified arbitrarily. Then the maximal function $M(\phi)$ becomes a function, $M(\phi_{\mathbf{H}}/\phi)$, of the single phase $\phi_{\mathbf{H}}$. It is then readily shown that, as a function of $\phi_{\mathbf{H}}$, $M(\phi_{\mathbf{H}}/\phi)$ has a unique maximum in the whole interval $(0, 2\pi)$ and the value of $\phi_{\mathbf{H}}$ which maximizes $M(\phi_{\mathbf{H}}/\phi)$ is given by the SAS tangent formula (3.6).

3.5 How to Find Solutions of the System of SAS Tangent Equations

Specify arbitrarily initial values for all the phases ϕ. Fix **H**. Calculate a new value for the phase $\phi_{\mathbf{H}}$ by means of the SAS tangent formula (3.6), in this way, in view of §3.4, increasing the initial value of the maximal function $M(\phi)$. Fix $\mathbf{H}' \neq \mathbf{H}$. Calculate a new value for $\phi_{\mathbf{H}'}$, using Eq. (3.6), the new value for $\phi_{\mathbf{H}}$, and initial values for the remaining phases, thus increasing still further the value of $M(\phi)$. Continue in this way to obtain new values for all the phases, thus completing the first iteration and, in the process, continuously increasing the value of $M(\phi)$. Complete as many iterations as necessary in order to secure convergence. Convergence is assured since the iterative process yields a monotonically increasing sequence of numbers, the values of $M(\phi)$, bounded above by unity. Evidently the process leads to a solution of the system of tangent equations (3.4) and, at the same time, a local maximum of the maximal function $M(\phi)$.

The applications show that only rarely does convergence require more than thirty iterations, and usually fewer than twenty suffice.

3.6 The Tallest Peak, Totally Isolated, Smooth and Ripple Free, Rests on the Broadest Base

The process described in §3.5 always leads to a local maximum of $M(\phi)$, the number of which is legion. It is natural therefore to ask: Will the process ever lead to the global maximum? Preliminary calculations, based on

three structures ranging in complexity from 1000 to 4000 independent atoms and in resolution from 3.0 to 2.5Å, show unequivocally that the answer is yes, frequently! Quite unexpectedly, the success rate is high, usually in the range of 10 to 15%. Hence 1000 trials yield, as it turns out, at most two distinct global maxima having, however, almost identical values, at least 100 times in typical cases.

A remarkable additional feature of the SAS maximal function is the total isolation of its global maxima. Thus, while the values of the local maxima are continuously distributed in a rather narrow range, the values of the global maxima exceed these by far. This remarkable property of $M(\phi)$ not only makes it easy to identify its global maxima and the associated sets of values for the phases, but no doubt accounts as well for the unexpectedly large circle of convergence surrounding each global maximum.

3.7 The SAS Correspondence Principle

It is clear from §§3.4 and 3.5 that there corresponds to every solution of the system of SAS tangent equations (3.6) a local maximum of $M(\phi)$, and conversely (the SAS correspondence principle).

3.8 The Linear Congruence Connection

The problem of going from the estimated values $\omega_{\mathbf{HK}}$ of the three-phase structure invariants ϕ_{HK} (2.5) to the values of the individual phases ϕ may be formulated as the problem of solving the redundant system of linear congruences

$$\phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}} \equiv \omega_{\mathbf{HK}} (\mathrm{mod\,ulo}\ 2\pi) \qquad (3.7)$$

each with weight $A_{\mathbf{HK}}$. This was the point of view adopted in the earlier work of Han *et al.* [7] and Hauptman and Han [8] in which the system (3.7) was transformed into a redundant system of linear equations

$$\phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}} = \omega_{\mathbf{HK}} + 2\pi n_{\mathbf{HK}} \qquad (3.8)$$

and the attempt was made, with limited success, to determine the integers $n_{\mathbf{HK}}$ in such a way as to make the system (3.8) self consistent. The resulting redundant system of linear equations (3.8) was then solved by least-squares.

The SAS maximal principle may thus be re-interpreted as yielding the solution of the redundant system of linear congruences (3.7). Somewhat unexpectedly, in those cases where the SAS maximal function has two global maxima, the system of redundant *linear* congruences (3.7) has two solutions, only one of which is the proper solution of the phase problem.

3.9 The Initial Application

The method described here was applied, with experimentally determined diffraction data, to the *ab initio* solution of the phase problem for the platinum derivative of the previously known macromomycin structure [9] consisting of approximately 750 protein atoms and 150 solvent molecules and crystallizing in the space group $P2_1$. With diffraction data to 2.5Å resolution, 150,000 three-phase structure invariants with largest A values were estimated. These involved 2710 phases whose values were to be determined.

One hundred solutions of the system of SAS tangent equations (3.6) were obtained using initial values of the phases chosen at random. Each of these trials converged to solution in five to eight cycles. Of the 100 trials, 17 yielded the same global maximum of the SAS maximal function $M(\phi)$ (3.5) which, in this case, turned out to be unique. Since the macromomycin structure had been previously determined, it was possible to calculate the average initial phase error using the known phases from the refined structure. This turned out to be 49˚ for all 2710 phases. It should be stressed that this solution of the phase problem for macromomycin was strictly *ab initio* in the sense that the only information needed were the observed SAS diffraction intensities at 2.5Å resolution; and the resulting map was interpretable.

With error-free diffraction intensities, the same calculation, again using SAS estimates for 150,000 three-phase structure invariants, yielded the values of 2120 phases with an initial average phase error of 30˚.

Research supported by the National Institutes of Health Program Project Grant No. GM46733.

References

[1] H. Hauptman, "A Minimal Principle in the Phase Problem", in *Crystallographic Computing 5 from Chemistry to Biology*; Proceedings of the International School of Crystallographic Computing", Bischenberg, France (1990); D. Moras, A.D. Podjarny & J.C. Thierry (Eds.), pp. 324-332. IUCr Oxford University Press, 1991.

[2] H. Hauptman, D. Velmurugan, & H. Fusen, "The Minimal Principle Solves Some Crystal Structures", in *Direct Methods of Solving Crystal Structures* (H. Schenk, Ed.), Proceedings of the International School of Crystallography, Erice, Trapani, Italy, April (1990), Plenum Publ., New York, New York, pp. 403-406, 1991.

[3] G. DeTitta, C. Weeks, P. Thuman, R. Miller, & H.Hauptman, Structure Solution by Minimal Function Phase Refinement and Fourier Filtering. I. Theoretical Basis, Acta Cryst., A50, 203-210, 1994.

[4] C.M. Weeks, G.T. DeTitta, H.A. Hauptman, P. Thuman, R.Miller, Structure Solution by Minimal Function Phase Refinement and Fourier Filtering. II. Implementation and Applications, ActaCryst., A50, 210-220, 1994.

[5] R. Miller, S.M. Gallo, H.G. Khalak, & C.M. Weeks, SnB: Crystal Structure Determination via Shake-and-Bake, J. Appl. Cryst., 27, 613-621, 1994.
[6] H. Hauptman, On Integrating the Techniques of Direct Methods with Anomalous Dispersion: I. The Theoretical Basis, Acta Cryst. A38, 632-641 1982.
[7] F. Han, G. DeTitta, & H. Hauptman, TELS: Least-Squares Solution of the Structure Invariant Equations, Acta Cryst. A47, 484-490, 1991.
[8] H. Hauptman & F. Han, Phasing Macromolecular Structures Via Structure Invariant Algebra, Acta Cryst., D49, 3-8, 1993.
[9] P. Van Roey & T.A. Beerman, Crystal Structure Analysis of Auromomycin Apoprotein (Macromomycin) Shows Importance of Protein Sidechains to Chromophore Binding Selectivity, Proc. Natl. Acad. Sci.-USA, 86, 6587-6591, 1989.