The Bayesian Statistical Approach to the Phase Problem in Protein X-ray Crystallography

Zhijun Wu, George Phillips, Richard Tapia, and Yin Zhang

CRPC-TR99791
April 1999

Center for Research on Parallel Computation
Rice University
6100 South Main Street
CRPC - MS 41
Houston, TX 77005

Submitted November 1999; also available as TR99-13,
Department of Computational and Applied Mathematics,
Rice University

The Bayesian Statistical Approach to the Phase Problem in Protein X-ray Crystallography*

Zhijun Wu[†], George Phillips[‡], Richard Tapia[§], and Yin Zhang[¶]

Abstract. We review a Bayesian statistical approach to the phase problem in protein X-ray crystallography. We discuss the mathematical foundations and the computational issues. The introduction to the theory and the algorithms does not require strong background in X-ray crystallography and related physical disciplines.

Key Words. X-ray crystallography, protein structure determination, Bayesian statistical theory, entropy maximization, maximum likelihood calculation, nonlinear optimization, fast Fourier transform

1 Introduction

X-ray crystallography is the most practical approach to protein structure determination. In a total of about 8,000 protein structures deposited in the

*Technical Report, TR99-13, Department of Computational and Applied Mathematics, Rice University, April 1999

[†]Department of Biochemistry and Cell Biology, and Keck Center for Computational Biology, Rice University, Houston, Texas, 77005

[‡]Department of Computational and Applied Mathematics, and Keck Center for Computational Biology, Rice University, Houston, Texas, 77005

[§]Department of Computational and Applied Mathematics, and Keck Center for Computational Biology, Rice University, Houston, Texas, 77005

[¶]Department of Computational and Applied Mathematics, and Keck Center for Computational Biology, Rice University, Houston, Texas, 77005. This author was supported in part by DOE Grant DE-FG03-97ER25331.

BNL PDB data bank, more than 80 percent of them were obtained through X-ray crystallography. With the development of recombinant DNA technology, which makes protein crystallization and amino acid sequence determination more feasible than ever before, even more structures are expected to be solved through X-ray crystallography in the next five to ten years.

The basis of crystallography is to determine the structure for a protein through the X-ray diffraction pattern produced by crystals of the protein. To prepare for this, the protein needs to be purified, crystallized, and examined by special equipment to obtain a set of X-ray patterns. Once these procedures are completed, the diffraction data are examined and the structure is deduced [39, 15, 23, 24].

An X-ray beam, when passing through a protein crystal, is diffracted through scattering by the electrons of the atoms in the protein. The diffracted X-ray is therefore related to the distribution of the electrons in the crystal. The diffraction pattern can be specified by a set of complex numbers called structure factors, $\{F_{\mathbf{H}} : \ F_{\mathbf{H}} = |F_{\mathbf{H}}|e^{i\alpha_{\mathbf{H}}}\}$, each reflecting in some sense the brightness of the X-ray light (the amplitude $|F_{\mathbf{H}}|$) and the light wave's origin (the phase $\alpha_{\mathbf{H}}$). Let the electron density distribution of a crystal system be denoted by a function $\rho(\mathbf{r})$ where $\mathbf{r}$ is a three dimensional vector representing an arbitrary point in the three dimensional space. Then, there is a correlation between the electron density distribution and the structure factors by the following Fourier transform pair,

$$F_{\mathbf{H}} = \int_{\mathbf{V}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{H} \cdot \mathbf{r}) \ \mathbf{dr}, \tag{1}$$

$$\rho(\mathbf{r}) = \sum_{\mathbf{H}} F_{\mathbf{H}} \exp(-2\pi i \mathbf{H} \cdot \mathbf{r}), \tag{2}$$

where $\mathbf{V}$ is the unit cell of the crystal and $\mathbf{H}$ a three-dimensional integer vector.

¿From equations (1) and (2), we see that if we know all the structure factors, we can immediately obtain the electron density distribution function, and vice versa. Once we have the electron density distribution function, a standard procedure can be used to obtain the atomic coordinates via analysis of electron density contouring. However, in practice, we do not fully know the structure factors from the experimental data. ¿From the X-ray diffraction measurements, we see only the magnitudes, while the phases are missing. Here, there arises the well-known phase problem, which has chal-

lenged scientists for decades to find an efficient and reliable solution to it, that is, given all the intensities (or amplitudes) of the structure factors, find all the phases and then use them to determine the structure of the crystal.

If we view the crystal as a set of separate atoms each surrounded by its electrons, the equation in (1) can be simplified to the following form,

$$F_{\mathbf{H}} = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \tag{3}$$

where n is the number of atoms in the unit cell of the crystal, $\mathbf{r}_j$ is the three dimensional position of atom j, and f_j is called the atomic scattering factor,

$$f_j = \int_{\mathbf{V}_j} \rho(\mathbf{r}) \exp(2\pi i \mathbf{H} \cdot \mathbf{r}) \ \mathbf{dr}, \tag{4}$$

where $\mathbf{V}_j \subset \mathbf{V}$ is the volume containing only atom j. It is reasonable to assume that all atoms of a given type have the same electron density distribution. Then, the integral in (4) can be calculated and tabulated with quantum mechanical methods for every different atomic type.

Still, given the amplitudes, there can be arbitrarily many possible values for the phases. Therefore, the phase problem seems not fully defined and has even been considered unsolvable in the past. However, if we write $F_{\mathbf{H}}$ in (3) in the explicit complex form, we will be able to obtain a set of nonlinear equations,

$$|F_{\mathbf{H}}| \exp(i\Phi_{\mathbf{H}}) = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \tag{5}$$

for all $\mathbf{H}$ for which $|F_{\mathbf{H}}|$ is given in the diffraction data. Let m be the number of structure factors, and consider both the real and imaginary parts of each equation. We can then have total $2m$ equations, while there are only $3n + m$ unknowns, the coordinates $\mathbf{r}_j$ and the phases $\alpha_{\mathbf{H}}$. Since usually $m \gg n$ for small molecules, the unknowns are over-determined, and in principle, the equations can be solved. In particular, a solution can be obtained by minimizing a nonlinear least squares function for the equations subject to a set of constraints that agree with certain structural invariant and semi-invariant properties of the phases. Hauptman and Karle [28] developed a method based on these observations and applied it successfully to small and

centrosymmetric systems. Their work followed by later developments led them, in particular one of them, Herbert Hauptman, as a mathematician, to win the Nobel prize in chemistry in 1984.

The equations in (5) become difficult to solve when applied to large molecules such as proteins. First, it is difficult to obtain enough diffraction data for large systems as required. Second, which is also related to the first, the solution for such a system usually converges slowly, if not diverges, due to inherent approximations, large search spaces, as well as poor initial points. Several research groups have worked on improving the algorithms so that they can be applied to large molecules, such as with the Shake-and-Bake method by Hauptman, et al [10, 11, 26].

In this paper, we review an alternative approach to the phase problem, the Bayesian statistical approach, proposed and pursued by Bricogne and several others in the last ten years or so [1, 2, 6, 3, 4, 5, 12, 13, 14, 17, 20, 21, 22, 41, 45]. This approach finds a solution to the phase problem by using statistical techniques rather than satisfying the equations in (5). With this approach, inaccuracy in the data or the model are better tolerated, different levels of experimental knowledge can be incorporated in the solution procedure, and the solution is deduced with Bayesian statistical inference tools.

The central idea of the Bayesian statistical approach is to use the Bayesian Theorem to evaluate the posterior probability of any hypothetical values for a subset of phases given the prior knowledge of the amplitudes for all the structure factors. A set of values for the phases are selected when the posterior probability is maximized. The process can be repeated by adding more trial phases until all of them are included and the crystal structure is determined.

In order to take this approach, the set of structure factors often is divided into two subsets, a basis set H and a non-basis set K. Algorithms are designed to select or refine the phases in the basis set, and extend them by including more phases from the non-basis set. The given condition is the known values for the amplitudes of the factors in both basis and non-basis sets, and the decision for accepting a set of phases in H is based on the conditional probability $P(H/K)$ for H given K, meaning the probability of having a crystal structure with factors in H given the fact that it already has the factors in K.

Recall in the statistical theory [35, 34], given a set of random events, C_j, $j = 1, \ldots, n$, with C_j's pairwise disjoint, and an additional event D, the

conditional probability of C_j for any j given D is computed by the Bayesian Theorem,

$$P(C_j/D) = \frac{P(C_j)P(D/C_j)}{P(D)}, \qquad P(D) = \sum_i P(C_i)P(D/C_i). \tag{6}$$

By using the Bayesian Theorem, the conditional probability $P(H/K)$, or more accurately, the posterior probability of H given K, should be

$$P(H/K) = \frac{P(H)P(K/H)}{P(K)}, \qquad P(K) = \sum_H P(H)P(K/H). \tag{7}$$

The Bayesian statistical approach to the phase problem requires evaluating, or in other words, maximizing the conditional probability in (7) in every step, and therefore, its major computational components are computing the probabilities required in (7), and in particular, $P(H)$ and $P(K/H)$. One reasonable way of calculating the probability $P(H)$ is by using the maximum entropy theory of statistical mechanics and information theory. The conditional probability $P(K/H)$, according to standard statistical theory, is the likelihood of H giving rise to K. The computation of this probability requires maximum likelihood calculations.

In the following sections, we will describe in greater detail the Bayesian statistical approach to the phase problem. Following closely its development by Bricogne et al, we discuss the mathematical foundations and the computational issues, and show how the approach can be applied to phase determination, and in particular, phase refinement and extension. Our goal is to understand the mathematical structure and the computational problems through a complete and accurate description of the approach. We also wish to provide a relatively self-contained review on the approach so that readers do not need to search the literature or to have background or experience in X-ray crystallography and related physical disciplines. We start with a brief introduction to protein X-ray crystallography in Section 2. The historical development of the direct methods for phase determination is reviewed in Section 3. A full description of the Bayesian statistical approach is given in Section 4. The computational problems for entropy maximization and maximum likelihood calculation are discussed in Sections 5 and 6. The phase determination procedure is demonstrated for phase refinement and extension in Section 7. Comments and remarks are made in Section 8 on issues yet to be addressed in the approach and possible future developments.

2 Protein X-ray Crystallography

Before taking an X-ray diffraction data, the protein needs to be purified and crystallized, usually requiring substantial laboratory efforts which do not necessarily succeed every time. Many protein crystals take weeks, months, or even years to grow. A crystal consists of atoms arranged in a certain pattern that repeats periodically in three dimensions. It diffracts X-rays, and produces regular diffraction patterns which can be recorded on X-ray detectors. Since proteins are large molecules and the diffraction tends to be weaker than that from small molecules, powerful X-ray sources, such as synchrotron radiation, often are required for protein crystals.

An individual X-ray pattern typically shows a two dimensional image with numerous diffraction spots nicely arranged in lattices. Typically one hundred or so of these images are taken while slowly rotating the crystal in the X-ray beam to yield a set of diffraction intensities mapped in three-dimensions. Once this data set is obtained, the crystal structure can be deduced from the positions and intensities of the spots since they correspond to the three dimensional arrangement of the atoms in the crystal. However, a nontrivial mathematical problem needs to be solved to achieve this.

More accurately speaking, the X-ray diffraction is generated by the electrons in the crystal. When an electron is hit by the X-ray beam, it "feels" the wave of the X-ray light and oscillates. It scatters the X-ray as it oscillates, the result of which is what we observe as diffraction. The electrons move quickly around the atoms. Their configuration can only be described by an electron density distribution function. The atoms can then be identified in the high-density regions of the function.

The total diffraction from an atom with many electrons, called atomic scattering factor and given in (4), is assumed the same for the same type of atoms. It can therefore be calculated with standard quantum mechanical methods and tabulated for multiple uses.

In the above sense, the atoms can be regarded as higher level units than electrons scattering the X-ray beams. The diffraction from the whole crystal is an aggregation of all atomic scattering. Each structure factor is simply the complex sum of the scattering factors contributed from all the atoms to the particular factor as shown in (3).

The periodic structure of the crystal is key to the X-ray diffraction. Crystals are built with repeating structures, called unit cells, aligned in a certain

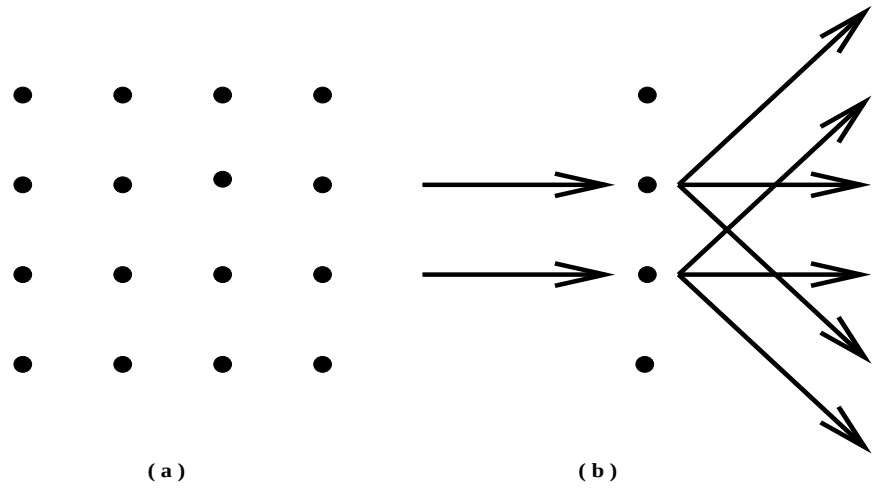

Figure 1: (a) A two-dimensional view of a crystal lattice, (b) X-rays scattered by the atoms in different directions

order in three dimensions. A unit cell may have one or more atoms, or one or more molecules, but all the cells have the same atoms or molecules and the same three-dimensional structure. If one atom appears in one unit cell, it appears with the same position in all others as well. All these atoms then form a three-dimensional lattice. The lattice structure is invariant with the choice of the origin and only depends on the structure of the crystal. Formally, a three-dimensional lattice is a set of points L defined as

$$L = \{\mathbf{p} \mid \mathbf{p} = \mu\mathbf{a} + \nu\mathbf{b} + \xi\mathbf{c},\ \mu, \nu, \xi \in Z,\ \mathbf{a}, \mathbf{b}, \mathbf{c} \in R^3\}, \tag{8}$$

where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are unit vectors specifying three lattice dimensions. For a crystal lattice, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ often are chosen for convenience to be the axes of the unit cell.

Suppose that we have a crystal lattice with an atom at each point of the lattice as shown in two dimensions in Figure 1. When atoms are hit by the X-ray beam, they start scattering X-rays in different directions. The diffracted X-rays from different atoms interfere and cancel each other except for some particular directions where they are enhanced instead. As a result, we observe certain patterns of diffraction from which the structural information can be derived.

X-rays are nothing more than very energetic light waves and can be described by a complex function,

$$X = A \exp(i\frac{2\pi\Phi}{\lambda}), \tag{9}$$

where X is the displacement of the wave, A the magnitude, λ the wavelength, and Φ the phase.

As shown in Figure 2, let A_1 and A_2 be two atoms with a distance in between equal to d, X_1 and X_2 two incident X-rays on A_1 and A_2, and X_1' and X_2' two scattered X-rays by A_1 and A_2, respectively. The scattered X-rays X_1' and X_2' have the same wavelength as the original X-rays, but their phases are different by an amount related to the distance $p + q$. ¿From Figure 2 we can see that

$$p + q = d \sin\theta_1 + d \sin\theta_2. \tag{10}$$

For X-rays X_1' and X_2' to be enhanced rather than canceled by each other, the phase difference in between has to be a multiple of the wavelength, that is,

$$d \sin\theta_1 + d \sin\theta_2 = n\lambda, \tag{11}$$

where n is an integer.

The equation in (11) is called the Laue Equation. Note that in general the angles θ_1 and θ_2 are not equal. However, the scattered X-rays X_1' and X_2' can be viewed as if they were reflected from X_1 and X_2 on two imagined reflecting planes R_1 and R_2, respectively, with a reflecting angle θ. Based on this observation, the Laue Equation can then be simplified as

$$2d' \sin\theta = n\lambda, \tag{12}$$

where d' is the distance between the two reflecting planes. The equation in (12) is called the Bragg's Law of X-ray diffraction, and X-ray diffraction is hence also called X-ray "reflection".

Now consider the atoms in the three-dimensional crystal lattice along directions **a**, **b**, and **c**. In order for a reflecting plane to have reflections enhanced from the atoms along all these directions, it must satisfy the Bragg's

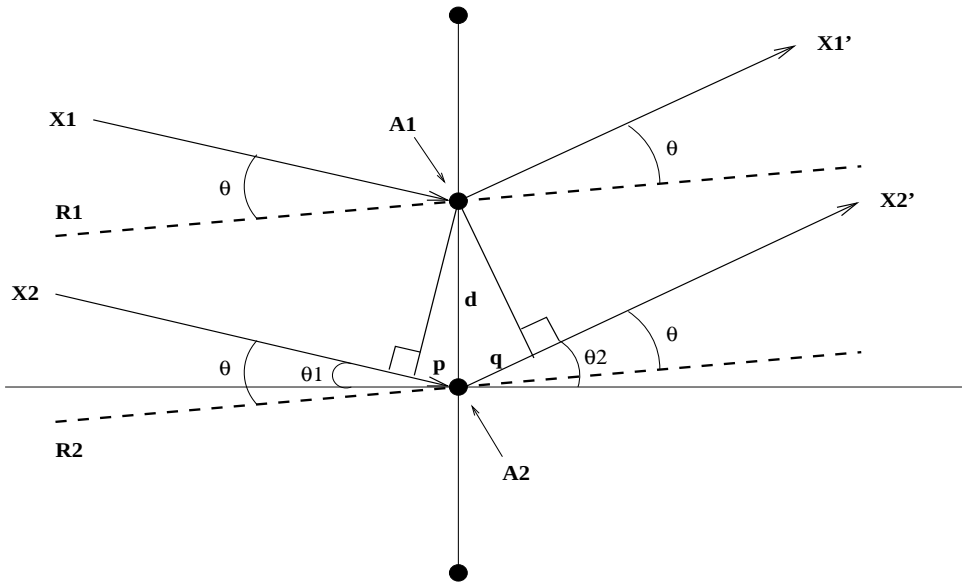

Figure 2: X-ray diffraction-reflection from a crystal lattice

Law for the atomic pairs along **a**, **b**, and **c** simultaneously. Therefore,

$$2d'_a \sin\theta = h\lambda, \tag{13}$$
$$2d'_b \sin\theta = k\lambda, \tag{14}$$
$$2d'_c \sin\theta = l\lambda, \tag{15}$$

where h, k, and l are integers, and d'_a, d'_b, and d'_c are the distances between the neighboring reflecting planes along **a**, **b**, and **c**, respectively. Since each of the reflecting planes corresponds to one of the reflections and is determined uniquely by the triplet (h, k, l), the latter, called the Miller index, is used as a label for the reflecting plane as well as the corresponding reflection. It also defines a point in the reciprocal lattice of the crystal.

Let a crystal lattice L be defined by (8). Then, the reciprocal lattice of L is defined as

$$L^* = \{\mathbf{p}^* \mid \mathbf{p}^* = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*,\ h,k,l \in Z\}, \tag{16}$$

where

$$\mathbf{a}^* = \frac{\mathbf{b}\times\mathbf{c}}{V},\ \mathbf{b}^* = \frac{\mathbf{a}\times\mathbf{c}}{V},\ \mathbf{c}^* = \frac{\mathbf{a}\times\mathbf{b}}{V}, \tag{17}$$

where V is the volume of the unit cell of the crystal lattice.

There are several important relationships between the crystal lattice and its reciprocal lattice. We state them in the following propositions.

Proposition 2.1 *Let* $\mathbf{a}$, $\mathbf{b}$, *and* $\mathbf{c}$ *be the unit vectors for the crystal lattice, and* $\mathbf{a}^*$, $\mathbf{b}^*$, *and* $\mathbf{c}^*$ *those for the reciprocal lattice. Then,*

$$\mathbf{a} \cdot \mathbf{a}^* = \mathbf{b} \cdot \mathbf{b}^* = \mathbf{c} \cdot \mathbf{c}^* = 1, \tag{18}$$

$$\mathbf{a} \cdot \mathbf{b}^* = \mathbf{a} \cdot \mathbf{c}^* = \mathbf{b} \cdot \mathbf{a}^* = \mathbf{b} \cdot \mathbf{c}^* = \mathbf{c} \cdot \mathbf{a}^* = \mathbf{c} \cdot \mathbf{b}^* = 0. \tag{19}$$

Proof. It follows immediately from the definition of $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$, and the facts that

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = V \tag{20}$$

$$\mathbf{b} \cdot (\mathbf{a} \times \mathbf{c}) = V \tag{21}$$

$$\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = V, \tag{22}$$

and

$$\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{a} \cdot (\mathbf{a} \times \mathbf{c}) = 0 \tag{23}$$

$$\mathbf{b} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{a} \times \mathbf{b}) = 0 \tag{24}$$

$$\mathbf{c} \cdot (\mathbf{a} \times \mathbf{c}) = \mathbf{c} \cdot (\mathbf{b} \times \mathbf{c}) = 0. \tag{25}$$

□

Proposition 2.2 *Let* $\mathbf{H}$ *be a vector, or in other words, a point in the reciprocal lattice,*

$$\mathbf{H} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*. \tag{26}$$

Then, H is normal to the reflecting plane (h, k, l) *in the crystal lattice.*

Proof. Let $\mathbf{r}$ be a vector on the reflecting plane (h, k, l) and

$$\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c}. \tag{27}$$

Let the projections of a distance vector $\mathbf{d}$ between two neighboring planes on $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ be d'_a, d'_b, and d'_c. Then, $\mathbf{d} = d'_a\mathbf{a} + d'_b\mathbf{b} + d'_c\mathbf{c}$, and d'_a, d'_b, and d'_c are proportional to h, k, and l, respectively, that is,

$$\frac{d'_a}{h} = \frac{d'_b}{k} = \frac{d'_c}{l} = \alpha \tag{28}$$

for some $\alpha \neq 0$.

Since the distance vector is perpendicular to the reflecting plane,

$$\mathbf{d} \cdot \mathbf{r} = d'_a x + d'_b y + d'_c z = \alpha(hx + ky + lz) = 0. \tag{29}$$

It follows that

$$\mathbf{H} \cdot \mathbf{r} = hx + ky + lz = 0, \tag{30}$$

and $\mathbf{H}$ is normal to the reflecting plane (h, k, l). □

Proposition 2.3 *Let A be an atom in the crystal lattice with a position $\mathbf{r}_A = x_A\mathbf{a} + y_A\mathbf{b} + z_A\mathbf{c}$. Let $\mathbf{H}$ be the norm of the reflecting plane (h, k, l), $\mathbf{H} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$. Then the contribution to the diffraction spot (h, k, l) from atom A, denoted $F^A_{\mathbf{H}}$, is a function of $\mathbf{r}_A$,*

$$F^A_{\mathbf{H}} = f_A \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_A) = f_A \exp[2\pi i(hx_A + ky_A + lz_A)], \tag{31}$$

where f_A is the atomic scattering factor of atom A.

Proof. By the definition of $F^A_{\mathbf{H}}$,

$$F^A_{\mathbf{H}} = \int_{\mathbf{V}_A} \rho(\mathbf{r}) \exp(2\pi i \mathbf{H} \cdot \mathbf{r}) d\mathbf{r}, \tag{32}$$

where $\mathbf{V}_A \subset \mathbf{V}$ is a volume containing only atom A. Let $\mathbf{r} = \mathbf{r}' + \mathbf{r}_A$. Then,

$$\begin{aligned} F^A_{\mathbf{H}} &= \int_{\mathbf{V}_A} \rho(\mathbf{r}' + \mathbf{r}_A) \exp(2\pi i \mathbf{H} \cdot \mathbf{r}') d\mathbf{r}' \ \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_A) && (33)\\ &= f_A \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_A) && (34)\\ &= f_A \exp[2\pi i(hx_A + ky_A + lz_A)]. && (35) \end{aligned}$$

□

The reflection from the whole unit cell of the crystal lattice is the sum of the reflections from all the atoms in the unit cell, and therefore,

$$F_{\mathbf{H}} = \sum_{j=1}^{n} F^j_{\mathbf{H}} = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j) = \sum_{j=1}^{n} f_j \exp[2\pi i(hx_j + ky_j + lz_j)], \tag{36}$$

where $F_{\mathbf{H}}$ is the reflection (h, k, l) from the whole unit cell, and is also called the structure factor (h, k, l) of the crystal lattice.

In typical X-ray crystallography experiments, each structure factor $F_{\mathbf{H}}$ corresponds to a spot of the light on the X-ray detector, and the intensity of the light is proportional to the square of the amplitude of $F_{\mathbf{H}}$. Therefore, the amplitudes of all $F_{\mathbf{H}}$ can be obtained by measuring the intensities of the recorded diffraction on the X-ray detector.

3 Direct Methods for the Phase Problem

The structure factor can be defined in a general form in terms of the electron density distribution function $\rho(\mathbf{r})$, that is,

$$F_{\mathbf{H}} = \int_{\mathbf{V}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{H} \cdot \mathbf{r}) \, \mathbf{dr}, \tag{37}$$

where $\mathbf{V}$ is the unit cell of the crystal. This formula reduces to (36) when the integration is taken by part for each of the atoms. On the other hand, $\rho(\mathbf{r})$ can be expanded as a Fourier series with the structure factors in (37) being the coefficients,

$$\rho(\mathbf{r}) = \sum_{\mathbf{H}} F_{\mathbf{H}} \exp(-2\pi i \mathbf{H} \cdot \mathbf{r}). \tag{38}$$

The formulas (37) and (38) show a direct relationship between the electron density function and the structure factors, or in other words, a relationship between the crystal structure and the X-ray diffraction pattern. If we know all the structure factors, we can immediately obtain the electron density distribution function and hence the atomic structure of the crystal. However, from X-ray diffraction experiments, all we know about the structure factors are their amplitudes but not the phases, given the fact that they are complex numbers and physically represent both intensities and directions of diffracted X-rays. In order to determine the crystal structure, we have to first solve a so-called phase problem, that is, find all the phases based on the given experimental data for the amplitudes so that the electron density distribution of the crystal can be fully determined.

The phase problem is not trivial to solve. Usually, not much knowledge is available about the structure of the crystal. Then, for a given set of values for the amplitudes, the phases can be anything to fulfill the equations in (37) and (38). On the other hand, if we set the phases to some arbitrary values, the function $\rho(\mathbf{r})$ they define will most likely give a meaningless structure. For example, it may become negative in some regions, which completely violates the physics.

However, if we consider the structure factor $F_{\mathbf{H}}$ as given in (36) and write it explicitly in its complex form, we then obtain the following equation,

$$|F_{\mathbf{H}}| \exp(i\Phi_{\mathbf{H}}) = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j). \tag{39}$$

Given f_j's *a priori* and $|F_{\mathbf{H}}|$ from the experiment, the equation has the unknowns $\Phi_{\mathbf{H}}$ and $\mathbf{r}_j$'s. For a whole set of $\mathbf{H}$, we obtain a system of equations over-determining the unknowns: If we separate the real and imaginary parts of the equations, we obtain $2m$ of them, where m is the number of known $|F_{\mathbf{H}}|$'s, while there are $m + 3n$ unknowns with m phases and $3n$ coordinates and $m \gg n$. Therefore, in principle, the phases together with the coordinates of the atoms can be determined by solving a system of equations defined by (39). The solution for the phases $\Phi_{\mathbf{H}}$ will provide a description of the crystal in terms of its electron density distribution. The values for the atomic coordinates $\mathbf{r}_j$ will suffice to determine the crystal structure at the atomic level, which should also be consistent with the results from the electron density distribution.

The phase problem has been a great challenge in X-ray crystallography. Many approaches, experimental, mathematical, and computational, have been pursued and applied to important applications. A direct method for the phase problem is a method of determining the phases for a given set of diffraction data by solving the equations (39) without assuming much prior knowledge about the structure. Historically, it also has a strong flavor of using probability theory to find the most probable solution to the phase problem. Work on direct methods started in 1950's. The most important landmark in the area was the book by Hauptman and Karle [28] published in 1953. Hauptman and Karle laid fundamental work for direct determination of the phases from the observed amplitudes. In particular, they studied equations (39) and developed a solution method in a probability-based least squares sense [27, 28, 31]. More specifically, phases and atomic positions in equations (39) were sought to fit the experimental data. In addition to the observed amplitudes, important relationships among phases or structure factors were also developed and used in the solution process, including the joint probability distribution of the structure factors, structural invariant and seminvariant properties of phases, etc. The work by Hauptman and Karle followed by later developments led to successful phase determination for small molecules, for example, with less than 100 atoms. Their method has been used in many of the laboratories as a routine procedure for crystal analysis.

Direct methods have been further studied and developed for decades since Hauptman and Karle. Improvements were made for more reliable and efficient procedures. However, a big challenge still remains for the methods to

apply for large molecules such as proteins, as demanded in the fast growing area of structural molecular biology. For large molecules, the data obtained from the X-ray experiment is not as sufficient and accurate as required by the direct methods. The system of equations (39) becomes difficult to solve because of too many variables (more than thousands). More information about the phase relationships is also necessary to restrict the search to more probable regions. Several new approaches have been pursued in recent years, including the convolution equation method , the maximum determinant method, as well as the original Hauptman-Karle method [18, 43, 25, 33, 40, 36, 44]. In particular, a so-called Shake-and-Bake method extended from the original Hauptman and Karle approach has been developed and tested on a set of small protein problems [10, 11, 26].

The direct methods have not been as successful for protein structure determination. Among all protein crystals studied so far, most of them were determined by using other more experimental approaches such as isomorphous replacement, molecular replacement, multiple wavelength anomalous dispersion, etc. [23, 24]. These methods incorporate more experimental or physical knowledge about the crystal structure and are able to determine specific molecules effectively. However, they are more problem specific, rely on the availability of structural knowledge or special experimental facilities, and therefore, cannot be applied to general cases. A general and efficient direct method for *ab initio* phase determination is still of great interest and importance to the structural determination of large molecules.

4 The Bayesian Statistical Approach

The Bayesian statistical approach uses statistical theory to derive the phases directly from the known amplitudes of the structure factors. The general principle applies to non-direct determination of the phases as well.

The central idea of this approach is using the Bayesian Theorem to evaluate the posterior probability of a set of structure factors given any prior knowledge of their amplitudes (or phases). The most probable ones are selected, from which the correct phases are determined.

The whole process starts from a small set of factors H,

$$H = \{F^*_{\mathbf{H}_j} : j = 1, \ldots, m_H\}. \tag{40}$$

Let K be another set of factors with all but H factors,

$$K = \{F^*_{\mathbf{K}_j} : j = 1, \ldots, m_K\}. \tag{41}$$

Then, the Bayesian Theorem is applied to compute $P(H/K)$ for any particular H given K, that is, the probability of having the factors in H given the factors in K. An optimal set H is selected when $P(H/K)$ is maximized. The next step is to expand H to include more factors from K. The process then repeats until H is expanded to the whole set of factors.

By using the Bayesian Theorem,

$$P(H/K) = \frac{P(H)P(K/H)}{P(K)}, \qquad P(K) = \sum_H P(H)P(K/H). \tag{42}$$

This formula reveals a relationship between the two sets of structure factors. It can be used for dual purposes, that is, if K is somehow fixed, H can be determined by solving an optimization problem,

$$\max_H \ P(H/K), \tag{43}$$

and if H is fixed, K can be optimized by

$$\max_K \ P(H/K). \tag{44}$$

Since the amplitudes for H and K are all known. The problems are essentially for determining the phases. The first problem is used for phase refinement or improvement for H and the second for extending phases to K with no previous phase information.

We now consider phase refinement. Note that $P(K)$ is constant no matter what H is. Therefore, the problem becomes

$$\max_H \ P(H)P(K/H). \tag{45}$$

According to the standard statistical theory, $P(K/H)$ is the likelihood of H giving rise to K denoted $\Lambda(H/K)$. So the problem in (45) really is to maximize the product of the probability of H and the likelihood of H giving rise to K. The first issue here is how to compute the probability and the likelihood. The probability $P(H)$ can be computed by considering the entropy of

the corresponding physical system, while $P(K/H)$ through maximum likelihood calculation. We explain the basic ideas in the following and leave more details in Section 5 and 6.

In order to compute the probability $P(H)$, we consider the entropy of the crystal system that produces H. The term entropy is from statistical physics and information theory [29, 30, 42, 37, 38]. It measures the uncertainty of a physical system or the amount of information a communication system may convey. Mathematically, if the movement of the particle in a physical system is described by a probability distribution function $q(\mathbf{r})$, where $\mathbf{r}$ is a random position of the particle, the entropy of the physical system then is defined as

$$\mathcal{H}(q) = -\int_{\mathbf{V}} q(\mathbf{r}) \log q(\mathbf{r}) \mathbf{dr}, \tag{46}$$

where $\mathbf{V}$ is the space that contains the system, and q is positive everywhere and

$$\int_{\mathbf{V}} q(\mathbf{r}) \mathbf{dr} = 1. \tag{47}$$

For two probability distributions q and p, the relative entropy of q with respect to p is defined as

$$\mathcal{S}_p(q) = -\int_{\mathbf{V}} q(\mathbf{r}) \log[q(\mathbf{r})/p(\mathbf{r})] \mathbf{dr}, \tag{48}$$

where p is also called a prior distribution of q.

It is easy to verify that a uniformly distributed system has the maximum entropy, that is, $\mathcal{H}_{max} = \mathcal{H}(\bar{m}) = \log V$, where $\bar{m}(\mathbf{r}) = 1/V$ is the uniform distribution function, and V is the volume of the system. We now consider the relative entropy of any q with respect to $\bar{m}$. From the definition in (48),

$$\mathcal{S}_{\bar{m}}(q) = -\int_{\mathbf{V}} q(\mathbf{r}) \log[q(\mathbf{r})/\bar{m}(\mathbf{r})] \mathbf{dr}. \tag{49}$$

Note $\mathcal{S}_{\bar{m}}(q)$ measures the entropy loss of a system. It is zero when q is equal to $\bar{m}$, and negative otherwise (losing entropy). On the other hand, if $\mathcal{S}_{\bar{m}}(q) = 0$, q must have the maximum uncertainty and be equal to the uniform distribution $\bar{m}$. If $\mathcal{S}_{\bar{m}}(q) < 0$, q must be restricted and less uniform, and $\mathcal{S}_{\bar{m}}(q)$ is the maximum entropy allowed by the restraints,

$$\max_q \quad \mathcal{S}_{\bar{m}}(q) \tag{50}$$

$$s.t. \quad q \in \mathcal{C}, \tag{51}$$

where $\mathcal{C}$ is the feasible set of q.

For a crystal system with a normalized electron density distribution $\rho(\mathbf{r})$, the entropy of the system can be calculated by solving problem (50. In particular, if ρ is such that the structure factors have the values in H, the entropy of ρ can be obtained by maximizing the entropy function subject to the constraints for the structure factors in H,

$$\max_\rho \quad \mathcal{S}_{\bar{m}}(\rho) \tag{52}$$

$$s.t. \quad \int_{\mathbf{V}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{H}_j \cdot \mathbf{r}) \mathbf{dr} = F^*_{\mathbf{H}_j}, \quad j = 1, \ldots, m_H \tag{53}$$

$$\int_{\mathbf{V}} \rho(\mathbf{r}) \mathbf{dr} = 1. \tag{54}$$

Let the maximum entropy be denoted by $\mathcal{S}^*_{\bar{m}}$ and the corresponding density distribution by ρ_H. Then,

$$\mathcal{S}^*_{\bar{m}} = \mathcal{S}_{\bar{m}}(\rho_H) = -\int_{\mathbf{V}} \rho_H(\mathbf{r}) \log[\rho_H(\mathbf{r})/\bar{m}(\mathbf{r})] \mathbf{dr}. \tag{55}$$

If the system has N electrons, the probability for the system to have a density distribution ρ_H and hence the structure factors in H can now be computed by the following formula,

$$P(H) = P(\rho_H) = \exp[N\mathcal{S}_{\bar{m}}(\rho_H)] = \exp(N\mathcal{S}^*_{\bar{m}}). \tag{56}$$

In Section 5, we will see how $\mathcal{S}^*_{\bar{m}}$ can be obtained by solving the entropy maximization problem (52) as a convex programming problem.

We now consider how the probability $P(K/H)$ is computed. The entropy idea for $P(H)$ also applies to $P(K/H)$ for which the prior distribution becomes ρ_H instead of $\bar{m}$. The entropy of any probability distribution ρ given the prior ρ_H is

$$\mathcal{S}_{\rho_H}(\rho) = -\int_{\mathbf{V}} \rho(\mathbf{r}) \log[\rho(\mathbf{r})/\rho_H(\mathbf{r})] \mathbf{dr}, \tag{57}$$

while the entropy of ρ with any constraints is calculated by

$$\max_\rho \quad \mathcal{S}_{\rho_H}(\rho) \tag{58}$$

$$s.t. \quad \rho \in \mathcal{C}, \tag{59}$$

where $\mathcal{C}$ is the feasible set of ρ. If $\mathcal{C}$ is defined such that ρ has the structure factors in K, the entropy of ρ can be computed by solving a similar problem as in (52),

$$\begin{aligned} \max_\rho \quad & \mathcal{S}_{\rho_H}(\rho) && (60)\\ s.t. \quad & \int_{\mathbf{V}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{K}_j \cdot \mathbf{r}) \mathbf{dr} = F^*_{\mathbf{K}_j}, \quad j = 1, \ldots, m_K && (61)\\ & \int_{\mathbf{V}} \rho(\mathbf{r}) \mathbf{dr} = 1. && (62) \end{aligned}$$

Let the maximum entropy be denoted by $\mathcal{S}^*_{\rho_H}$ and the corresponding density distribution by ρ_K. Then,

$$\mathcal{S}^*_{\rho_H} = \mathcal{S}_{\rho_H}(\rho_K) = -\int_{\mathbf{V}} \rho_K(\mathbf{r}) \log[\rho_K(\mathbf{r})/\rho_H(\mathbf{r})] \mathbf{dr}. \tag{63}$$

If the system has N electrons, the probability for the system to have a density distribution ρ_K and hence the structure factors in K given those in H can now be computed by the following formula,

$$P(K/H) = P(\rho_K) = \exp[N\mathcal{S}_{\rho_H}(\rho_K)] = \exp(N\mathcal{S}^*_{\rho_H}). \tag{64}$$

However, in practice, $F^*_{\mathbf{K}_j}$ are not fully given. Only their amplitudes are known. The constraints in (61) then become nonlinear, making the problem more difficult to solve. In Section 6 we describe how an approximate problem can be formulated and solved instead.

5 Entropy Maximization

We now consider how to solve the entropy maximization problem in (52). For convenience, we write the problem in the following general form,

$$\begin{aligned} \max_\rho \quad & \mathcal{S}_{\bar{m}}(\rho) && (65)\\ s.t. \quad & \mathcal{C}_j(\rho) = c_j = F^*_{\mathbf{H}_j}, \quad j = 1, \ldots, m && (66)\\ & \mathcal{C}_0(\rho) = c_0 = 1, && (67) \end{aligned}$$

where $\mathcal{C}_j$ are linear constraint functionals defined as

$$\begin{aligned} \mathcal{C}_j(\rho) \quad &= \quad \int_{\mathbf{V}} \rho(\mathbf{r}) C_j(\mathbf{r}) \mathbf{dr}, \quad j = 0, \ldots, m, && (68)\\ & && (69) \end{aligned}$$

with

$$C_j(\mathbf{r}) = \exp(2\pi i \mathbf{H}_j \cdot \mathbf{r}), \quad j = 1, \ldots, m \tag{70}$$

$$C_0(\mathbf{r}) = 1. \tag{71}$$

Since the objective function is concave and the constraints are linear, the problem is a convex program. In fact, the objective function is even strictly concave, and therefore, the solution to the problem must also be unique.

We now form the Lagrangian function for the problem as follows,

$$\mathcal{L}(\rho, \lambda_0, \ldots, \lambda_m) = \mathcal{S}_{\bar{m}}(\rho) + \sum_{j=0}^{m} \lambda_j [\mathcal{C}_j(\rho) - c_j]. \tag{72}$$

If ρ is a local maximizer of problem (65), the partial derivative of the Lagrangian function with respective to ρ is necessarily equal to zero. We then obtain,

$$-1 - \log[\rho(\mathbf{r})/\bar{m}(\mathbf{r})] + \sum_{j=0}^{m} \lambda_j C_j(\mathbf{r}) = 0. \tag{73}$$

Solve the equation for ρ to obtain

$$\rho(\mathbf{r}) = \bar{m}(\mathbf{r}) \exp(\lambda_0 - 1) \exp[\sum_{j=0}^{m} \lambda_j C_j(\mathbf{r})]. \tag{74}$$

Let $\lambda_0 - 1 = -\log Z$. Then,

$$\rho(\mathbf{r}) = \frac{\bar{m}(\mathbf{r})}{Z} \exp[\sum_{j=1}^{m} \lambda_j C_j(\mathbf{r})]. \tag{75}$$

Since ρ satisfies the normalization constraint (67),

$$\mathcal{C}_0(\rho) = \int_{\mathbf{V}} \rho(\mathbf{r}) C_0(\mathbf{r}) \mathbf{dr} = \int_{\mathbf{V}} \rho(\mathbf{r}) \mathbf{dr} = 1, \tag{76}$$

we then obtain Z as a function of $\lambda_1, \ldots, \lambda_m$,

$$Z(\lambda_1, \ldots, \lambda_m) = \int_{\mathbf{V}} \bar{m}(\mathbf{r}) \exp[\sum_{j=1}^{m} \lambda_j C_j(\mathbf{r})] \mathbf{dr}. \tag{77}$$

By applying other constraints (66) to ρ, we have

$$\int_{\mathbf{V}} \frac{\bar{m}(\mathbf{r})}{Z(\lambda_1,\ldots,\lambda_m)} \exp[\sum_{l=1}^{m} \lambda_l C_l(\mathbf{r})] C_j(\mathbf{r}) \mathbf{dr} = c_j, \tag{78}$$

for $j = 1,\ldots,m$. These equations can be used to determine $\lambda_1,\ldots,\lambda_m$, and hence ρ in terms of (75). A compact form of the equations can be written as

$$\partial/\partial_{\lambda_j}(\log Z)(\lambda_1,\ldots,\lambda_m) = c_j, \quad j = 1,\ldots,m. \tag{79}$$

We state these results formally in the following propositions.

Proposition 5.1 *Let ρ be a local maximizer of problem (65). Then there exist a set of parameters $\lambda_0,\ldots,\lambda_m$ such that*

$$\mathcal{S}'_{\bar{m}}(\rho) + \sum_{j=0}^{m} \lambda_j \mathcal{C}'_j(\rho) = 0, \tag{80}$$

$$\mathcal{C}_j(\rho) = \int_{\mathbf{V}} \rho(\mathbf{r}) C_j(\mathbf{r}) \mathbf{dr} = c_j, \quad j = 0,\ldots,m. \tag{81}$$

Proof. Given the fact that C_j and hence $\mathcal{C}'_j$ are linear independent, the regularity condition holds at ρ. Then, there must exist parameters $\lambda_0,\ldots,\lambda_m$ such that the first order necessary condition for ρ to be a local maximizer of (65) is satisfied, which implies that (80) and (81) are necessarily true. Moreover, since (65) is a convex program, the conditions are also sufficient. □

Proposition 5.2 *A set of parameters $\lambda_0,\ldots,\lambda_m$ satisfy the equations in (80) and (81) if and only if the parameters $\lambda_1,\ldots,\lambda_m$ solve the equations,*

$$\nabla(\log Z)(\lambda_1,\ldots,\lambda_m) = \mathbf{c}, \tag{82}$$

where $\mathbf{c} = (c_1,\ldots,c_m)^T$ and Z is defined as in (77).

Proof. The proof is as discussed in the beginning of the section and demonstrated through the derivation from (72) to (79). □

Let G be a function and $\langle G \rangle$ the average value of G by a probability distribution ρ,

$$\langle G \rangle = \int_{\mathbf{V}} \rho(\mathbf{r}) G(\mathbf{r}) \mathbf{dr}. \tag{83}$$

Then it is easy to verify that

$$\begin{aligned} \partial_j(\log Z) &= \langle C_j \rangle, & (84)\\ \partial^2_{jk}(\log Z) &= \langle C_j \overline{C}_k \rangle - \langle C_j \rangle \langle \overline{C}_k \rangle = \langle (C_j - \langle C_j \rangle)\overline{(C_k - \langle C_k \rangle)} \rangle, & (85) \end{aligned}$$

where $\overline{(C_k - \langle C_k \rangle)}$ is the complex conjugate of $(C_k - \langle C_k \rangle)$. It implies that the Hessian of $\log Z$, or the Jacobian of the entropy equations (82), is a covariance matrix of the deviation of C_j's from their averaged values.

Proposition 5.3 *The Hessian of* $\log Z$ *is the covariance matrix of the deviation of* C_j*'s from their averaged values by the probability distribution* ρ*, and*

$$\nabla^2(\log Z) = \langle (\mathbf{C} - \langle \mathbf{C} \rangle)(\mathbf{C} - \langle \mathbf{C} \rangle)^H \rangle, \tag{86}$$

where $\mathbf{C} = (C_1, \ldots, C_m)^T$, $(\mathbf{C} - \langle \mathbf{C} \rangle)^H$ *is the complex conjugate of* $(\mathbf{C} - \langle \mathbf{C} \rangle)$*, and* $\langle \rangle$ *is taken component-wise.*

Proof. By the definition of Z in (77),

$$\begin{aligned} \partial_j(\log Z) &= \frac{1}{Z}\partial_j Z & (87)\\ &= \int_{\mathbf{V}} \frac{\bar{m}(r)}{Z(\lambda_1 \ldots, \lambda_m)} \exp[\sum_{l=1}^{m} \lambda_l C_l(r)] C_j(r) dr & (88)\\ &= \int_{\mathbf{V}} \rho(r) C_j(r) dr = \langle C_j \rangle. & (89) \end{aligned}$$

It follows that

$$\begin{aligned} \partial^2_{jk}(\log Z) &= \frac{1}{Z}\partial^2_{jk} Z - \frac{1}{Z^2}\partial_j Z \partial_k Z & (90)\\ &= \langle C_j \overline{C}_k \rangle - \langle C_j \rangle \langle \overline{C}_k \rangle & (91)\\ &= \langle (C_j - \langle C_j \rangle)\overline{(C_k - \langle C_k \rangle)} \rangle. & (92) \end{aligned}$$

The Hessian of $\log Z$ is then obtained in the form of (86). □

Corollary 5.1 *The Hessian of* $\log Z$ *is positive definite.*

Proof. Let $\mathbf{x} = (x_1, \ldots, x_m)^T$ be a nonzero vector and $\mathbf{x}^H$ the complex conjugate of $\mathbf{x}$.

$$\begin{aligned} \mathbf{x}^H \nabla^2 (\log Z) \mathbf{x} &= \mathbf{x}^H \langle (\mathbf{C} - \langle \mathbf{C} \rangle)(\mathbf{C} - \langle \mathbf{C} \rangle)^H \rangle \mathbf{x} && (93) \\ &= \langle \mathbf{x}^H (\mathbf{C} - \langle \mathbf{C} \rangle)(\mathbf{C} - \langle \mathbf{C} \rangle)^H \mathbf{x} \rangle && (94) \\ &= \langle |\mathbf{x}^H (\mathbf{C} - \langle \mathbf{C} \rangle)|^2 \rangle && (95) \\ &\geq 0. && (96) \end{aligned}$$

Assume that the equality holds for some $\mathbf{x}$,

$$\mathbf{x}^H \nabla^2 (\log Z) \mathbf{x} = \langle |\mathbf{x}^H (\mathbf{C} - \langle \mathbf{C} \rangle)|^2 \rangle = 0. \tag{97}$$

We then have

$$\mathbf{x}^H (\mathbf{C} - \langle \mathbf{C} \rangle) = 0. \tag{98}$$

Given the fact that $C_j \neq \langle C_j \rangle$ and $C_j - \langle C_j \rangle$ are linear independent of each other, $\mathbf{x}$ must be equal to zero, contradicting to the assumption that $\mathbf{x}$ be a nonzero vector. Therefore,

$$\mathbf{x}^H \nabla^2 (\log Z) \mathbf{x} = \langle |\mathbf{x}^H (\mathbf{C} - \langle \mathbf{C} \rangle)|^2 \rangle > 0, \tag{99}$$

and $\nabla^2 (\log Z)$ is positive definite. □

Note that $\langle C_j \rangle$ corresponds to the structure factor $F_{\mathbf{H}_j}$ for some reciprocal vector $\mathbf{H}_j$, and $\langle C_j \overline{C}_k \rangle$ corresponds to $F_{\mathbf{H}_j - \mathbf{H}_k}$. Therefore,

$$\begin{aligned} \partial_j (\log Z) &= F_{\mathbf{H}_j}, && (100) \\ \partial^2_{jk} (\log Z) &= F_{\mathbf{H}_j - \mathbf{H}_k} - F_{\mathbf{H}_j} F_{-\mathbf{H}_k}. && (101) \end{aligned}$$

Since all $F_{\mathbf{H}_j}$ can be computed once in $\mathcal{O}(m \log m)$ calculations with fast Fourier transform, the gradient and Hessian of $\log Z$ can be assembled in $\mathcal{O}(m \log m)$ computation time.

Finally, we show that solving the maximum entropy equation (82) is equivalent to solving the dual problem of the maximization problem (65). According to the standard theory of convex programming [19, 16], the dual problem of (65) is a minimization problem for the Lagrangian function subject to a necessary condition that the partial derivative of the Lagrangian

function with respect to ρ is equal to zero, that is,

$$\min_{\rho,\lambda_0,\ldots,\lambda_m} \quad \mathcal{S}_{\bar{m}}(\rho) + \sum_{j=0}^{m} \lambda_j [\mathcal{C}_j(\rho) - c_j] \tag{102}$$

$$s.t. \qquad \mathcal{S}'_{\bar{m}}(\rho) + \sum_{j=0}^{m} \lambda_j \mathcal{C}'_j(\rho) = 0. \tag{103}$$

Solving ρ in the constraint and replace it in the objective function, we obtain the following equivalent unconstrained minimization problem,

$$\min_{\lambda_1,\ldots,\lambda_m} \quad \log Z - \sum_{j=1}^{m} \lambda_j c_j, \tag{104}$$

where Z is defined in the same way as in (77).

A necessary condition for $\lambda_1, \ldots, \lambda_m$ to be a solution to problem (104) is that the gradient of the objective function at $\lambda_1, \ldots, \lambda_m$ is equal to zero, and therefore,

$$\partial/\partial\lambda_j(\log Z) = c_j, \quad j = 1, \ldots, m., \tag{105}$$

which are the same entropy maximization equations as (82). Since the Hessian of the objective function is equal to $\nabla^2(\log Z)$ which is positive definite, the necessary condition (105) is also sufficient, and it determines $\lambda_1, \ldots, \lambda_m$ uniquely.

The problem (104) can be solved by a standard Newton's method, as proposed by Bricogne [1]. Let $\lambda = (\lambda_1, \ldots, \lambda_m)^T$. Then the Newton iteration for the problem can be formulated as follows.

$$\lambda^{(l+1)} = \lambda^{(l)} - \alpha^{(l)} [\nabla^2(\log Z)(\lambda^{(l)})]^{-1} [\nabla(\log Z)(\lambda^{(l)}) - \mathbf{c}], \tag{106}$$

where $\alpha^{(l)}$ is a step length. Since $\nabla^2(\log Z)(\lambda^{(l)})$ is always positive definite, the Newton's direction is descent at any point. With a line search procedure, the method will be able to decrease the function value in every step. If the function is bounded below, which is the case for problem (104), the method will eventually converge to the minimum. Moreover, it converges quadratically when the iterate is close to the optimal solution [9].

6 Maximum Likelihood Calculation

The conditional probability $P(K/H)$, or the likelihood $\Lambda(H/K)$, can be computed in a similar way as for $P(H)$ with entropy maximization. However, the constraints on the structure factors in K often are nonlinear, making the entropy maximization problem more difficult to solve: The factors are constrained to have their amplitudes equal to some given values. Then they cannot be solved for ρ explicitly. Also, the problem cannot necessarily be solved by solving a dual problem since the constraints are no longer convex.

An alternative way to compute $P(K/H)$ is to construct or approximate the probability distribution in terms of the structure factors. With electron density distribution,

$$P(K/H) = \exp[N\mathcal{S}_{\rho_H}(\rho_K)]. \tag{107}$$

Let $\mathbf{F}_{\rho_H} = (F_{\mathbf{H}_1}, \ldots, F_{\mathbf{H}_m})^T$ be the vector of the structure factors of ρ_H, and $\mathbf{F}_{\rho_K} = (F_{\mathbf{K}_1}, \ldots, F_{\mathbf{K}_m})^T$ of ρ_K. Then $\mathcal{S}_{\rho_H}(\rho_K)$ and hence $P(K/H)$ can also be defined as a function in terms of F_{ρ_H} and F_{ρ_K}. Let the function be denoted by $\mathcal{S}_{\mathbf{F}_{\rho_H}}(\mathbf{F}_{\rho_K})$. The conditional probability $P(K/H)$ then becomes

$$P(K/H) = \exp[N\mathcal{S}_{\mathbf{F}_{\rho_H}}(\mathbf{F}_{\rho_K})]. \tag{108}$$

By definition, $\mathcal{S}_{\rho_H}(\rho_K)$ has the property that it is equal to zero as $\rho_K = \rho_H$ and decreases to negative infinity as ρ_K deviates from ρ_H. Correspondingly, $\mathcal{S}_{\mathbf{F}_{\rho_H}}(\mathbf{F}_{\rho_K})$ should also be equal to zero as $\mathbf{F}_{\rho_K} = \mathbf{F}_{\rho_H}$ and decrease to negative infinity as $\mathbf{F}_{\rho_K}$ deviates from $\mathbf{F}_{\rho_H}$. The probability $P(K/H)$ can therefore be approximated by a Gaussian distribution function,

$$P(K/H) \approx \exp[-N(\mathbf{F}_{\rho_K} - \mathbf{F}_{\rho_H})^H \mathbf{Q}_H^{-1}(\mathbf{F}_{\rho_K} - \mathbf{F}_{\rho_H})], \tag{109}$$

where $\mathbf{Q}_H$ is the covariance matrix of the structure factors in $\mathbf{F}_{\rho_H}$, which by (100) is actually the Hessian of $\log Z$, and can be obtained as a by-product of the entropy maximization of $P(H)$.

The joint probability distribution of a set of structure factors can be expanded as an Edgeworth series [32], which can be approximated asymptotically as a Gaussian distribution function centered at the origin. The distribution (109) can also be viewed as such an asymptotic approximation, where the Gaussian is centered at $\mathbf{F}_{\rho_H}$ to define a conditional probability

distribution of the structure factors. While it has the same asymptotic property as the previous one, this approximation can especially be used for regions away from the origin as well.

The formula (109) can be used to calculate the conditional probability of K given any H, or in other words, the likelihood of H giving rise to K. The likelihood is a function of H and K. If H is fixed, the likelihood can be maximized by varying K, and vice versa. In the former case, the amplitudes of K usually are given, and therefore, the likelihood is maximized when a set of optimal phases for K are selected. For simplicity, let $\mathbf{F}_{\rho_K}$ be denoted by $\mathbf{F}$, $\mathbf{F}_{\rho_H}$ by $\mathbf{F}_0$, and $\mathbf{Q}_H$ by $\mathbf{Q}_0$. Then the problem can be formulated as,

$$\max_{\mathbf{F}} \quad \exp[-N(\mathbf{F}-\mathbf{F}_0)^H \mathbf{Q}_0^{-1}(\mathbf{F}-\mathbf{F}_0)], \tag{110}$$

$$s.t. \quad |F_{\mathbf{K}_j}| = |F^*_{\mathbf{K}_j}|, \quad j = 1, \ldots, m_K. \tag{111}$$

Note that in (110), the objective function is convex, but the constraints are nonlinear and non-convex. Therefore, the problem may have multiple maxima, and a global optimization algorithm may be required to find the maximum likelihood. Bricogne [1] and Bricogne and Gilmore [6] described several solution methods for the problem, all related to specific applications. A general algorithm for the problem is yet to be developed.

7 Phase Refinement and Extension

The Bayesian statistical approach to the phase problem can be used in various contexts of phase determination, direct and non-direct, whenever statistical inference is required to derive phases from partially available knowledge. Bricogne [5] described some of such applications as in molecular replacement, multi-wavelength anomalous dispersion, phase refinement and extension, structure refinement, etc. We will use phase refinement and extension as examples to show how the general approach is applied. The potential for developing a general direct phase determination algorithm is discussed.

A simple phase refinement process is described in the outline **Phase Refinement**. First, a set of structure factors H is specified for which the phases are to be refined. A set of values are then assigned to the phases, and the probabilities, $P(H)$ and $P(K/H)$, are calculated by the methods described in previous sections. This process is repeated for different sets of phase values until the maximum of the product of the two probabilities is reached and the

Phase Refinement

1. Initialize phases in H;
2. Compute $P(H)$ and obtain $\mathbf{Q}_H$;
3. Construct Gaussian approximation to $P(K/H)$;
4. Compute $P(K/H)$ given all amplitudes in K;
5. If the maximum of $P(H)P(K/H)$ is reached, stop;
6. Assign new values to the phases in H, and go to 2;

phases are refined with a set of optimal values. Note that the amplitudes of the factors in both H and K are known, and the phases in H are assigned to given values and are varied in the process, but the phases in K are unknown. So in order to compute $P(K/H)$, the phases in K need to be fixed or integrated out from the distribution function. Bricogne and Gilmore [6] approximated the covariance matrix $\mathbf{Q}_H$ by its diagonal elements. The phases in K can then be easily integrated out, and the distribution function becomes depending on only the phases in H together with the given amplitudes for all the factors.

Phase extension is referred to as a process to determine some unknown phases based on a given set of phases as well as the amplitudes for all the factors. For example, we may already have all low resolution phases, but we want to extend them to include all high resolution ones as well. A typical procedure for phase extension is outlined in **Phase Extension**. Suppose that a set of structure factors H is given and all phases in H are known. Let K be the set of structure factors whose phases are to be determined. First we need to compute the probability $P(H)$ and obtain a covariance matrix $\mathbf{Q}_H$. We can then construct a Gaussian distribution function as an approximation to the probability distribution of $P(K/H)$. The phases for the structure factors in K can be determined by maximizing $P(K/H)$ subject to the constraints that the amplitudes of these factors must be equal to the known values. Once

Phase Extension

1. Compute $P(H)$ and obtain $\mathbf{Q}_H$;
2. Construct Gaussian approximation to $P(K/H)$;
3. Maximize $P(K/H)$ to obtain an optimal K;
4. Extend H to include K;

an optimal set of phases are determined, they can be included in H, and H is said to be extended to K.

Phase refinement and extension can be combined to construct a general phase determination algorithm. First we start from a small set of structure factors H, and apply a refinement procedure to H to obtain a set of phases for H. We then choose a small set of structure factors from K, and extend H to this set of factors. The whole process can be repeated for the extended H until all structure factors are included in H. Note that in the extension process, multiple minima may be obtained for the likelihood maximization problem. Therefore, H may be extended to several possible sets of factors, each subject to further expansion. The entire procedure will then proceed as spanning a tree of structure factors. Hopefully, an optimal set of factors can be found at the end of the tree branches. The whole procedure is outlined in **Iterative Refinement and Extension**. For more detailed descriptions about the refinement, extension, as well as full determination of the phases using the Bayesian statistical approach, readers are referred to all the references by Bricogne et al.

8 Remarks

Focusing on the phase problem, this paper introduces the theory and practice in protein X-ray crystallography. In particular, a Bayesian statistical approach to the phase problem is reviewed. The mathematical and computational issues in this approach are discussed. Two of the major compo-

Iterative Refinement and Extension

1. Put initial H in $\mathcal{T}$;
2. Select a H from $\mathcal{T}$;
3. Refine the phases in H;
4. Extend H to $H_1, \ldots, H_l$;
5. Set $\mathcal{T} = \{H_1, \ldots, H_l\} \cup \mathcal{T} \setminus H$;
6. If $\mathcal{T}$ is empty, stop; otherwise, go to 2;

nents, entropy maximization and maximum likelihood calculation, are described along with their formulations, mathematical properties, and solution methods. General algorithms for phase refinement, extension, and full-scale determination are presented and discussed. The paper is intended to introduce computational X-ray crystallography to computer scientists and applied mathematicians. Therefore, principles of X-ray crystallography, historical development of direct methods for the phase problem, as well as the fundamental theory of the Bayesian statistical approach are all introduced in great detail. Proofs for some of the related mathematical results are also provided. The goal of the paper is to understand the fundamental problems and motivate cross-disciplinary interests among computer science, applied mathematics, and X-ray crystallography that may result in fruitful collaborations in solving the critical computational problems in protein X-ray crystallography.

We conclude the paper by discussing two computational issues in the Bayesian statistical approach to the phase problem, one related to solving the entropy maximization problem and the other to maximizing the likelihood.

As we have described in the paper, the entropy maximization problem for computing the probability $P(H)$ is a strictly convex program and can be solved efficiently by using a standard Newton's method. While this is true in the sense that the Newton's method converges to the solution in fewer iterations (local quadratic convergence) than other methods, it requires $\mathcal{O}(m_H^3)$

computation in each iteration, where m_H is the number of structure factors in H. This can be a computational bottleneck when H becomes large since the phase determination algorithm requires solving the entropy maximization problem many times, especially for phase refinement. In order to work around the problem, it is suggested in [1] to compute the Hessian of $\log Z$ approximately with $F_{\mathbf{H}_j-\mathbf{H}_k}$ instead of $F_{\mathbf{H}_j-\mathbf{H}_k} - F_{\mathbf{H}_j}F_{-\mathbf{H}_k}$ for every (j,k) element. The inverse of the Hessian can then be obtained through fast Fourier transform rather than numerical factorization, reducing the computation to $\mathcal{O}(m_H \log m_H)$. However, the convergence rate of the algorithm may be slowed down because of the approximation. The algorithm must be used only when the Newton's method becomes un-affordable.

As we have mentioned in the paper, the problem (110) for maximizing the likelihood function $P(K/H)$ may have multiple local solutions when the amplitudes of the structure factors in K are restricted to some given values. Although in practice the global maximum may not be necessary, when it is desired, the problem may become very difficult to solve. Note that when the probability distribution $P(K/H)$ is approximated by a Gaussian, the maximization problem is equivalent to minimizing the quadratic exponent subject to the amplitude constraints. If we reformulate the problem in real space, it can be written in the following general form,

$$\min_{\mathbf{x}} \quad (\mathbf{x}-\mathbf{x}_0)^T A_0 (\mathbf{x}-\mathbf{x}_0) \tag{112}$$

$$s.t. \quad x_{2j-1}^2 + x_{2j}^2 = b_j^2, \quad j = 1, \ldots, m_K, \tag{113}$$

where A_0 is equivalent to $\mathbf{Q}_0$ in (110), $\mathbf{x} = (x_1, \ldots, x_{2m_K})^T$, and x_{2j-1} and x_{2j} correspond to the real and imaginary parts of the jth component of $\mathbf{F}$. In general, this problem can be hard to solve. For example, for a centrosymmetric system, $x_{2j} = 0$ for all j, and the problem becomes a discrete optimization problem where x_{2j-1} can take only two discrete values. However, the problem can be reduced to an unconstrained optimization problem. For example, it is equivalent to,

$$\min_{\alpha_1,\ldots,\alpha_{m_K}} (\mathbf{x}(\alpha)-\mathbf{x}_0)^T A_0 (\mathbf{x}(\alpha)-\mathbf{x}_0), \tag{114}$$

where $\mathbf{x}(\alpha) = (x_1, \ldots, x_{2m_K})^T$, $\alpha = (\alpha_1, \ldots, \alpha_{m_K})^T$, $x_{2j-1} = b_j \cos\alpha_j$ and $x_{2j} = b_j \sin\alpha_j$. However, a global minimizer is still required for the problem while the objective function apparently has many local minimizers because of the trigonometric functions.

The computational issues discussed above are critical for efficient implementation of phase determination algorithms with the Bayesian statistical approach. Work on the issues is important for future development of the approach.

Acknowledgment

This paper is a product of the Computational Structural Biology Seminar held at Rice CAAM Department during the spring semester of 1999. All participants of the seminar are acknowledged.

References

[1] G. Bricogne, *Maximum Entropy and the Foundations of Direct Methods*, Acta Cryst. A40, 1984, pp. 410–445.

[2] G. Bricogne, *A Bayesian Statistical Theory of the Phase Problem. I. A Multichannel Maximum-Entropy Formalism for Constructing Generalized Joint Probability Distributions of Structure Factors*, Acta Cryst. A44, 1988, pp. 517–545.

[3] G. Bricogne, *A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. III. Extension to Powder Diffraction Data*, Acta Cryst. A47, 1991, pp. 803–829.

[4] G. Bricogne, *Direct Phase Determination by Entropy Maximization and Likelihood Ranking: Status Report and Perspectives*, Acta Cryst. D49, 1993, pp. 37–60.

[5] G. Bricogne, *Bayesian Statistical Viewpoint on Structure Determination: Basic Concepts and Examples*, in Methods in Enzymology, Vol. 276, 1997, Academic Press, pp. 361 – 423.

[6] G. Bricogne and C. J. Gilmore, *A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. I. Theory, Algorithms and Strategy*, Acta Cryst. A46, 1990, pp. 284–297.

[7] P. L. Britten and D. M. Collins, *Information Theory as a Basis for the Maximum Determinant*, Acta Cryst. A38, 1982, pp. 129–132.

[8] H. E. Daniels, *Saddlepoint Approximations in Statistics*, Ann. Math. Stat. 25, 1954, pp. 631–650.

[9] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1983.

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

[11] G. T. DeTitta, C. M. Weeks, P. Thuman, R. Miller, and H. Hauptman, *Structure Solution by Minimal-Function Phase Refinement and Fourier Filtering. II. Implementation and Applications*, Acta Cryst. A50, 1994, pp. 210–220.

[12] W. Dong, T. Baird, J. R. Fryer, C. J. Gilmore, D. D. MacNicol, G. Bricogne, D. J. Smith, M. A. O'Keefe, and S. Hövmoller, *Electron Microscopy at 1-ÅResolution by Entropy Maximization and Likelihood Ranking*, Nature Vol. 355, 1992, pp. 605–609.

[13] S. Doublie, S. Xiang, C. J. Gilmore, G. Bricogne, and C. W. Carter Jr., *Overcoming Non-Isomorphism by Phase Permutation and Likelihood Scoring: Solution of the TrpRS Crystal Structure*, Acta Cryst. A50, 1994, pp. 164–182.

[14] S. Doublie, G. Bricogne, C. J. Gilmore, and C. W. Carter Jr., *Tryptophanyl-tRNA Synthetase Crystal Structure Reveals an Unexpected Homology to Tyrosyl-tRNA Synthetase*, Structure 15, 1995, 3:17–31.

[15] J. Drenth, *Principles of Protein X-ray Crystallography*, Springer-Verlag, New York, New York, 1994.

[16] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, New York, New York, 1987.

[17] E.L. Fortelle and G. Bricogne, *Maximum-Likelihood Heavy-Atom Parameter Refinement for Multiple Isomorphous Replacememt and Multiwavelength Anomalous Diffraction Methods*, in Methods in Enzymology, Vol. 276, 1997, pp. 472 – 494.

[18] G. Germain and Woolfson, *On the Application of Phase Relationships to Complex Structures*, Acta. Cryst. B24, 1968, pp. 91–96.

[19] P. E. Gill, W. Murray, M. H. Wright, *Practical Optimization*, Academic Press, Inc., New York, New York, 1981.

[20] C. J. Gilmore, G. Bricogne, and C. Bannister, *A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. II. Applications to Small Molecules*, Acta Cryst. A46, 1990, pp. 297–308.

[21] C. J. Gilmore, K. Henderson, and G. Bricogne, *A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. IV. The Ab Initio Solution of Crystal Structures from Their X-ray Powder Data*, Acta Cryst. A47, 1991, pp. 830–841.

[22] C. J. Gilmore, K. Henderson, and G. Bricogne, *A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. V. The Use of Likelihood as a Discriminator of Phase Sets Produced by the SAYTAN Program for a Small Protein*, Acta Cryst. A47, 1991, pp. 842–846.

[23] J. P. Glusker and K. N. Trueblood, *Crystal Structure Analysis*, Oxford University Press, New York, New York, 1985.

[24] J. P. Glusker, M. Lewis, and M. Rossi, *Crystal Structure Analysis for Chemists and Biologists*, VCH Publishers, Inc., New York, New York, 1994.

[25] H. Hauptman, *On the Identity and Estimation of Those Cosine Invariants, $Cos\ (\phi_m + \phi_n + \phi_p + \phi_q)$, Which Are Probably Negative*, Acta Cryst. A30, 1974, pp. 472–476.

[26] H. HAUPTMAN, *A Minimal Principle in the Phase Problem of X-ray Crystallography*, in Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos, D. Shalloway, and G. Xue, Eds., American Mathematical Society, 1996, pp. 89–96.

[27] H. HAUPTMAN AND J. KARLE, *Crystal-Structure Determination by Means of a Statistical Distribution of Interatomic Vectors*, Acta Cryst. 5, 1952, pp. 48–59.

[28] H. HAUPTMAN AND J. KARLE, *Solution of the Phase Problem I. The Centrosymmetric Crystal*, American Crystallography Association Monograph No. 3, Polycrystal Book Service, Pittsburgh, Pennsylvania, 1953.

[29] E. T. JAYNES, *Information Theory and Statistical Mechanics*, Phys. Rev. Vol. 106, pp. 620–630.

[30] E. T. JAYNES, *Prior Probabilities*, IEEE Trans. SSC-4, pp. 227–241.

[31] J. KARLE AND I. L. KARLE, *The Symbolic Addition Procedure for Phase Determination for Centrosymmetric and Noncentrosymmetric Crystals*, Acta Cryst. 21, 1966, pp. 849–859.

[32] A. KLUG, *Joint Probability Distributions of Structure Factors and the Phase Problem*, Acta Cryst. 11, 1958, pp. 515–543.

[33] M. F. C. LADD AND R. A. PALMER, *Theory and Practice of Direct Methods in Crystallography*, Plenum Press, New York, New York, 1980.

[34] P. M. LEE, *Bayesian Statistics: An Introduction*, Oxford University Press, New York, New York, 1989.

[35] B. W. LINDGREN, *Statistical Theory*, Macmillan Publishing Co., Inc., New York, New York, 1976.

[36] R. NARAYAN AND R. NITYANANDA, *The Maximum Determinant Method and the Maximum Entropy Method*, Acta Cryst. A38, 1982, pp. 122–128.

[37] O. PENROSE, *Foundations of Statistical Mechanics*, Pergamon Press, New York, New York, 1970.

[38] O. E. Piro, *Information Theory and the 'Phase Problem' in Crystallography*, Acta Cryst. A39, 1983, pp. 61–68.

[39] D. E. Sands, *Introduction to Crystallography*, Dover Publications, Inc., New York, New York, 1975.

[40] D. Sayre, *Computational Crystallography*, Oxford University Press, New York, New York, 1982.

[41] K. Shankland, C. J. Gilmore, G. Bricogne, and Hashizume, *A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. VI. Automatic Likelihood Analysis via the Student t Test, with an Application to the Powder Structure of Magnesium Boron Nitride,* Mg_3BN_3, Acta Cryst. A49, 1993, pp. 493–501.

[42] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, Illinois, 1963.

[43] G. Tsoucaris, *A New Method for Phase Determination. The 'Maximum Determinant Rule'*, Acta Cryst. A26, 1970, pp. 492–499.

[44] S. W. Wilkins, J. N. Varghese, and M. S. Lehmann, *Statistical Geometry. I. A Self-Consistent Approach to the Crystallographic Inversion Problem Based on Information Theory*, Acta Cryst. A39, 1983, pp. 47–60.

[45] S. Xiang, C. W. Carter Jr., G. Bricogne, and C. J. Gilmore, *Entropy Maximization Constrained by Solvent Flatness: a New Method for Macromolecular Phase Extension and Map Improvement*, Acta Cryst. D49, 1993, pp. 193–212.