Phase retrieval in crystallography and optics

R. P. Millane
The Whistler Center for Carbohydrate Research, Smith Hall, Purdue University, West Lafayette, Indiana 47907

Received September 26, 1989; accepted November 20, 1989

Phase problems occur in many scientific disciplines, particularly those involving remote sensing using a wave field. Although there has been much interest in phase retrieval in optics and in imaging in general over the past decade, phase retrieval has a much longer history in x-ray crystallography, and a variety of powerful and practical techniques have been developed. The nature of crystallography means that crystallographic phase problems are distinct from those in other imaging contexts, but there are a number of commonalities. Here the principles of phase retrieval in crystallography are outlined and are compared and contrasted with phase retrieval in general imaging. Uniqueness results are discussed, but the emphasis is on phase-retrieval algorithms and areas in which results in one discipline have, and may, contribute to the other.

1. INTRODUCTION

In several scientific fields, particularly those involving imaging with a wave field (e.g., microscopy, radio engineering, astronomy, crystallography), it is necessary or convenient to measure the Fourier transform of a quantity of interest rather than the quantity itself. Further, it often occurs (if the wavelength of the radiation used is small or if the radiation passes through a distorting medium, for example) that it is difficult or impossible to measure the phase of the Fourier transform accurately. Since the phase must be known if one is to reconstruct the image straightforwardly by inverse Fourier transformation, this constitutes a phase problem.

Practical solutions to the phase problem were first studied in the context of x-ray crystallography in the 1940's and have since developed into powerful techniques. These particular methods apply only to small molecules. A different class of techniques, applicable to large molecules, was first developed in the mid 1950's. Practical solutions to the phase problem in contexts other than crystallography (I will call this general imaging) began at later dates. Although there was some discussion in the 1960's of the significance of the loss of phase, it was not until the late 1970's and early 1980's that practical methods were developed in general imaging contexts.

Phase-retrieval theory and algorithms in crystallography and in general imaging have developed fairly independently. This is probably due to fundamental differences between these two problems that have necessitated different approaches. However, there are a number of commonalities, and the object of this paper is to outline the principles of crystallographic phase retrieval to the optics community (noting the similarities and differences with optics), with a view to prompting a flow of ideas between the two disciplines.

In order to set the scene, the salient aspects of phase retrieval in general imaging are reviewed in Section 2. In Section 3 I outline the basic theory of x-ray crystallography in which the phase problem arises. The three different classes of crystallographic phase problem are defined in Section 4. In Sections 5-7 solutions to the phase problem in each of these classes are outlined. Crystallographic and optical phase retrieval are compared in Section 8, and concluding remarks are made in the final section.

2. PHASE PROBLEM IN GENERAL IMAGING

General-Purpose Results and Algorithms

Consider an image denoted by \( f(x) \), where \( x \) is the position vector in image space. Image space is usually two dimensional, although there are cases in which it may be one dimensional (e.g., signal recovery) or more than two dimensional (e.g., electron microscopy of thick specimens). The Fourier transform of the image, or the visibility, \( F(u) \), is given by

\[
F(u) = \int f(x) \exp(i2\pi xu) dx,
\]

where \( u \) is the position vector in Fourier space and the integral is, implicitly, over all the \( N \)-dimensional image space. The image is given by the inverse Fourier transform of the visibility, so that

\[
f(x) = \int F(u) \exp(-i2\pi xu) du,
\]

and recovery of the image is straightforward (in principle) if \( F(u) \) is measured. A phase problem arises when only the amplitude \( |F(u)| \) can be measured. Since the image is given uniquely by Eq. (2), recovery of the image is equivalent to recovering the phase \( \phi(u) \), where

\[
F(u) = |F(u)| \exp[i\phi(u)].
\]

As a result of the autocorrelation theorem, \( \phi(u) \) is recovered from the amplitude by

\[
A(x) = \int f(y)f^*(x+y) dy,
\]

can be reconstructed from the magnitude, since \( |F(u)|^2 \) is the Fourier transform of \( A(x) \).

There are some properties of the image that are irretrievably lost when the phase is lost. This is most easily seen by noting that

\[
f(x), f(x + \xi_1) \exp(i\phi_1), \text{ and } f^*(-x + \xi_2) \exp(i\phi_2)
\]

are not unique, but that

\[
f(x + \xi_1) \exp(i\phi_1) = f(x) \exp(i2\pi xu) dx,
\]

where \( u \) is the position vector in Fourier space and the integral is, implicitly, over all the \( N \)-dimensional image space. The image is given by the inverse Fourier transform of the visibility, so that

\[
f(x) = \int F(u) \exp(-i2\pi xu) du,
\]

and recovery of the image is straightforward (in principle) if \( F(u) \) is measured. A phase problem arises when only the amplitude \( |F(u)| \) can be measured. Since the image is given uniquely by Eq. (2), recovery of the image is equivalent to recovering the phase \( \phi(u) \), where

\[
F(u) = |F(u)| \exp[i\phi(u)].
\]

As a result of the autocorrelation theorem, \( \phi(u) \) is recovered from the amplitude by

\[
A(x) = \int f(y)f^*(x+y) dy,
\]

can be reconstructed from the magnitude, since \( |F(u)|^2 \) is the Fourier transform of \( A(x) \).

There are some properties of the image that are irretrievably lost when the phase is lost. This is most easily seen by noting that

\[
f(x), f(x + \xi_1) \exp(i\phi_1), \text{ and } f^*(-x + \xi_2) \exp(i\phi_2)
\]

are not unique, but that

\[
f(x + \xi_1) \exp(i\phi_1) = f(x) \exp(i2\pi xu) dx,
\]

The object of this paper is to outline the principles of crystallographic phase retrieval to the optics community (noting the similarities and differences with optics), with a view to prompting a flow of ideas between the two disciplines.

In order to set the scene, the salient aspects of phase retrieval in general imaging are reviewed in Section 2. In Section 3 I outline the basic theory of x-ray crystallography in which the phase problem arises. The three different classes of crystallographic phase problem are defined in Section 4. In Sections 5-7 solutions to the phase problem in each of these classes are outlined. Crystallographic and optical phase retrieval are compared in Section 8, and concluding remarks are made in the final section.
all have the same visibility amplitude \(|F(u)|\), where \(\xi_1, \xi_2, \phi_1\), and \(\phi_2\) are real constants and * denotes complex conjugation.\(^1\)

Hence the absolute position, inversion in the origin, and a constant phase factor of the image are irrevocably lost, although this is usually of little practical significance.

If the image is an arbitrary complex function, the phase problem has a multitude of solutions since one can associate any phase function \(\phi(u)\) with the measured \(|F(u)|\) to generate different images. O’Neill and Walther\(^2\) and Walther\(^3\) appear to have been the first to recognize that the problem posed in this way does not correspond to any real physical situation. In any practical case, there are constraints on the form of the image. For example, in an optical system the image has finite support (i.e., the image is nonzero only within a finite region of image space), is square integrable, and usually is nonnegative. In general, applying any phase function to the amplitude will produce an image that violates these conditions. The important uniqueness question is, therefore: Is there more than one phase function that gives an image that satisfies the given conditions? Walther\(^3\) showed that in the one-dimensional case there are multiple, but countable, phase functions that lead to images with the correct support. He pointed out that nonnegativity could be expected to restrict the number of solutions further, although to what degree was not clear. He also noted that extension of these results to two dimensions “will probably be fraught with difficulties.”

Nonuniqueness in one dimension can be demonstrated as follows. When the one-dimensional Fourier space coordinate \(u = u\) is generalized to the complex variable \(w = u + iw\), the visibility is given by

\[
F(w) = \int_{-\infty}^{\infty} f(x) \exp(2\pi i wx) dx,
\]

where \(x\) is the single coordinate in image space. Since the image has compact support and is square integrable, \(F(w)\) is an entire function (analytic in the finite complex plane). Hence \(F(w)\) can be written as the Hadamard product

\[
F(w) = \exp(A + Bw) \prod_{n=1}^{M-1} \left(1 - w/w_n\right),
\]

so that \(F(w)\) is completely characterized by its complex zeros \(w_n\) (and the constants \(A\) and \(B\)). Defining \(I(w) = F(w) F^*(w^*)\), which is the analytic continuation of \(|F(u)|^2\) since \(I(u) = |F(u)|^2\), gives

\[
I(w) = \exp[2 \text{Re}(A) + 2 \text{Re}(B)w] \prod_{n=1}^{M-1} \left(1 - w/w_n\right) \left(1 - w/w_n^*\right),
\]

where \(\text{Re}\) denotes the real part, so that the zeros of \(I(w)\) are \(w_n\) and \(w_n^*\). From \(I(w)\), obtained by analytic continuation of \(|F(u)|^2\), one cannot determine which of \(w_n\) or \(w_n^*\), for each pair of zeros, corresponds to a zero of \(F(u)\). For a continuous image, the zeros tend toward the real axis as \(|w| \to \infty\), so that only a finite number (related to the noise level) have significant imaginary parts. The complex-conjugate ambiguity therefore applies effectively to a finite number, \(M\), say, of the zeros. Hence there are \(2^{M-1}\) solutions to the phase problem, as interchanging all the zeros gives \(f^*(-x)\) rather than \(f(x)\). Each of the resulting images has the same support.\(^3\)

The quite different situation in two dimensions appears to have been noted first by Bruck and Sodin.\(^4\) In two (or more) dimensions the Hadamard factorization [Eq. (6)] is generally not possible, and the nonuniqueness argument evaporates. It appears therefore that the ambiguity, if any, is substantially reduced over that for the one-dimensional case. Support for this result was obtained by Fienup,\(^5\) using iterative two-dimensional phase-retrieval algorithms. He described as “dramatic” his observation that test images reconstructed uniquely, independently of the image used to start the iterative procedure.

Uniqueness of the phase problem in two dimensions for positive real images was intimated by Bates\(^6\) in 1982 and, more convincingly, in 1984,\(^7\) using arguments from sampling theory. Uniqueness for general complex two-dimensional images was shown by Barakat and Newsam\(^8\) in 1984 in the sense that “multiplicity of solutions is pathologically rare,” using the theory of functions of two complex variables. Uniqueness was demonstrated, in a more practical sense, by Lane \textit{et al}\(^9\) and by Izraelevitz and Lim,\(^10\) using the concept of zero sheets, and their arguments were incorporated into reconstruction algorithms. (These concepts have also been studied by other researchers in the context of multidimensional signal processing.\(^11\)\(^12\)) In two dimensions, the visibility \(F(u) = F(u_1, u_2)\) can be analytically continued into a space with two complex dimensions, denoted by \(F(w) = F(w_1, w_2) = F(u_1, u_2, v_1, v_2)\), where the notation is obvious. \(F(w)\) therefore exists and is analytic within this four-dimensional space. Considering \(w_1, v_1\) say, fixed, \(F(w_1, w_2)\) has point zeros in the \(w_2\) complex plane. As \(w_1\) is varied, these zeros migrate along smooth two-dimensional surfaces (zero sheets) in the four-dimensional space, since \(F(w)\) is analytic.\(^8\)\(^10\) The Hadamard product in two dimensions almost always has a single factor, so that these sheets actually form a single sheet. This means that the point zeros cannot be treated (and flipped) independently as in the one-dimensional case. Only in special cases (that occur with probability zero in practice) do the zeros form more than one sheet. Since \(F(w)\)\(^13\) has two factors, it contains two zero sheets, one corresponding to \(F(w)\) and one to \(F^*(w^*)\). Although two-dimensional surfaces in a four-dimensional space almost never intersect, these two surfaces, being interdependent, intersect wherever \(w = \bar{w}\), i.e., wherever one of them crosses the real axis. Since the visibility is an entire function, continuity of the first derivative can be applied to distinguish between the two zero sheets at these points. Algorithms have been devised and implemented\(^10\) to separate the zero sheets and reconstruct the image. However, these algorithms have been applied only to simulated data with small images, since they are sensitive to noise and computationally intensive. It is worth noting that the concept of zero sheets has led to profound results concerning blind deconvolution.\(^13\)

Although the results described above are useful in demonstrating uniqueness, almost all practical phase-recovery algorithms effective for general images in which there is no phase information and the data are contaminated with noise are based on Fienup’s modifications\(^14\) of the Gerchberg-Saxton algorithm.\(^15\) These algorithms start with a random image \(f_0(x)\) and update it in an iterative fashion. At the \(n\)th iteration, a modified image \(f_n(x)\) is calculated by cycling through Fourier space, using the procedure...
where $\mathcal{F}[\cdot]$ denotes the Fourier-transform operation and $\arg[\cdot]$ denotes the phase. The updated image $f_{n+1}(x)$ can be formed in different ways to formulate different algorithms, the most straightforward being error reduction, for which

$$f_{n+1}(x) = f_n(x) \quad \text{if } x \in \Omega$$

$$= 0 \quad \text{otherwise,}$$

(9)

where $\Omega$ denotes the set of points $x$ where $f'(x)$ satisfies the image constraints. The error-reduction algorithm often converges slowly, and the hybrid input–output algorithm, defined by

$$f_{n+1}(x) = f_n(x) + \gamma [f_n(x) - f_n'(x)] \quad \text{otherwise,}$$

(10)

where $\gamma$ is a constant (typically between 0.5 and 1.0), called the feedback parameter, is usually more effective. Various practical considerations have to be considered when one is using these algorithms, and the reader is referred to the relevant references. Phase-retrieval algorithms of this type can also be implemented by using generalized projections, and the reader is referred to Ref. 19 for details on this topic.

Positivity is usually an important constraint in the application of iterative methods. If the image is complex, then reconstruction is more difficult, and a much tighter support constraint is needed. This is discussed further in Section 8.

Another image-restoration technique that has been applied to phase retrieval is the maximum-entrophy method. This is a general method for producing an image from limited and noisy data. Basically, it is a model-fitting procedure that seeks to recover an image that is consistent with given constraints (extent, positivity, measured Fourier amplitude, etc.) while maximizing an entropy functional such as $-\int f(x) \ln[f(x)]dx$. The rationale for this method is that such an image contains minimal information consistent with the data. This method has been applied to phase retrieval and (particularly) to spectral extrapolation (super-resolution). Although there appear to have been some successful applications, when and how the method should be applied in specific situations is still poorly understood. The maximum-entrophy method is not discussed further here, and the reader is referred to Refs. 22–24 for more information.

Speckle Imaging

The phase-retrieval techniques described above are general purpose in the sense that they can be applied in any situation when the Fourier modulus is measured. One area in which phase retrieval has been intensively studied and utilized is optical (and radio) astronomical imaging, in which the phase is perturbed by propagation of the radiation through the turbulent atmosphere. The particular characteristics of this situation have resulted in a number of special techniques for estimating the phase, most of which fall into the class of speckle imaging. In view of the importance of these techniques, and because of some relationships to crystallographic techniques, their salient points are outlined here (see Refs. 25 and 26 for reviews).

If an object is viewed through the atmosphere (or any distorting medium) and the distortion introduced by the medium is severe, the resolution is limited by the medium rather than by the imaging instrument (telescope). However, Labeyrie showed that diffraction-limited (by the telescope) resolution (but not images) can be achieved by appropriate processing of short-exposure (~10 msec, i.e., less than the fluctuation time of the atmosphere) speckle images. The advantage of the presence of the fluctuating atmosphere is that one can obtain a set of these speckle images, denoted by $s_m(x)$, each of which is distorted by independent manifestations of the atmosphere but all obeying the same statistics. Labeyrie recognized that an ensemble average (denoted by $\langle \cdot \rangle$) of the intensities of visibilities $|S_m(u)|^2$ of a set of speckle images is related to the (diffraction-limited) visibility intensity $|F(u)|^2$ by

$$\langle |S_m(u)|^2 \rangle = |F(u)|^2 |\langle H_m(u) \rangle|^2,$$

(11)

where $|\langle H_m(u) \rangle|^2$ is called the speckle transfer function. The speckle transfer function can be estimated by performing the same processing when viewing an unresolvable (point) object, and $|F(u)|^2$ can be obtained from Eq. (11) by inverse filtering. This is called speckle interferometry and eliminates the decrease in resolution that is due to the turbulent atmosphere. The phase of $F(u)$ can then be retrieved by using Fienup’s algorithms. However, the speckle images themselves contain some phase information, and speckle imaging is concerned with using these images to reconstruct the object (or phase) directly. If the object contains two components, one of which is unresolvable, then, under some circumstances, the method of speckle holography can be used to reconstruct the image. The image can be written as

$$f(x) = \delta(x) + g(x),$$

(12)

where $\delta(x)$ is the Dirac delta function, so that the autocorrelation, determined from the visibility intensity, is given by

$$A(x) = \delta(x) + g(x) + g^*(-x) + g(x) \odot g(x),$$

(13)

where $\odot$ denotes correlation. If the separation between the reference $\delta(x)$ and the remaining image $g(x)$ is greater than the width (extent) of $g(x)$, then the terms in Eq. (13) are distinct, allowing $g(x)$, or $g^*(-x)$, to be identified. If the reference object is sufficiently brighter than the unknown object, then the autocorrelation term in Eq. (13) is much smaller than $g(x)$, and the separation condition can be relaxed to the reference not lying within the unknown object. A similar approach is described by Liu and Lohmann in which the object is made up of a number of distinct features that are separated from one another by at least twice the diameter of the seeing disk (the Airy disk for the atmosphere).

Another method for reconstructing the phases is described by Knox and Thompson, in which the cross correlation of each speckle image is computed so that

$$\langle S_m^*(u)S_m(u + \alpha) \rangle = F^*(u)F(u + \alpha) |\langle H_m^*(u)H_m(u + \alpha) \rangle|,$$

(14)

An estimate of $|\langle H_m^*(u)H_m(u + \alpha) \rangle|$ is obtained by performing the same processing on a point object, and $F^*(u)F(u + \alpha)$ is recovered by inverse filtering. For $u = 0$, this gives...
\[ F^*(0)F(\alpha), \] which gives the phase of \( F(\alpha) \) since the phase of \( F(0) \) is zero. Furthermore, for \( u = \alpha \) we have \( F^*(\alpha)F(2\alpha) \), which gives the phase of \( F(2\alpha) \). Hence \( \arg[F(n\alpha)] \) can be determined recursively for all integer \( n \). By choosing different values of \( n \) it is possible to build up the phases of \( F(u) \) on a grid in Fourier space. Using these phases with the amplitudes from speckle interferometry allows the image to be reconstructed. Bootstrapping the phases in this way is error sensitive but can be stabilized by using least-squares techniques. Optimum values for \( \alpha \) are chosen based on experience. Error sensitivity probably limits the application of this technique. The Knox–Thompson technique is somewhat akin to the method of closure phase used in radio astronomy.\(^{32}\) Lynds, Worden, and Harvey\(^{33}\) (LWH) realized that it might be possible to form a useful image by superimposing the brighter speckles within a speckle image. Their technique can be described as follows: The amplitudes \( a_{m\eta} \) and positions \( x_{m\eta} \) of the peaks in the \( m \)th speckle image that are above a threshold value are used to generate a speckle mask \( \mu_m(x) \), given by

\[ \mu_m(x) = \sum_j a_{m\eta} \delta(x - x_{m\eta}). \]  

The image is formed by correlating the \( m \)th speckle mask with the \( m \)th speckle image (this has the effect of superimposing the bright speckles) and averaging over the speckle images, giving

\[ f_{\text{LWH}}(x) = \langle \tilde{s}_m(x) \otimes \mu_m(x) \rangle. \]  

This technique is difficult to apply to weak images and works best for images containing a strong unresolved component.\(^{34}\) The shift-and-add (sa) method\(^{35}\) is a special case of the LWH approach more suitable for faint images in which one chooses only the brightest point (with position \( x_m \)) in each speckle image, shifts this point to the origin of image space, and then averages over the speckle images, giving

\[ f_{\text{sa}}(x) = \langle s_m(x - x_m) \rangle. \]  

A promising method of phase retrieval from speckle images is the triple-correlation or bispectrum method.\(^{34}\) This is also called the speckle-masking method, not to be confused with the use of a speckle mask described above. The triple correlation \( f(x_1, x_2) \) and the bispectrum \( F(u_1, u_2) \) are defined by

\[ f(x_1, x_2) = \int_{-a}^{a} f(x)f(x + x_1)f(x + x_2)dx \]  

and

\[ F(u_1, u_2) = F(u_1)F(u_2)F(-u_1 - u_2), \]  

respectively, where \( f(x_1, x_2) \) and \( F(u_1, u_2) \) form a Fourier-transform pair. In the context of speckle imaging, the speckle bispectrum \( S_m(u_1, u_2) \) is related to the image bispectrum by

\[ \langle S_m(u_1, u_2) \rangle = \langle F(u_1, u_2)F^*(u_1, u_2) \rangle, \]  

where \( \langle F(u_1, u_2) \rangle \) is the bispectral speckle transfer function that can be estimated by applying the same processing to an unresolvable object, allowing \( F(u_1, u_2) \) to be calculated by inverse filtering. The bispectral speckle transfer function is positive under fairly wide conditions\(^{36,37}\) so that the phase of the object bispectrum equals that of the image bispectrum. The remarkable property of the bispectrum is that the phase of the \( F(u) \) can be calculated recursively from the phase of \( F(u_1, u_2) \). This is illustrated in the sampled one-dimensional case, for which Eq. (19) becomes

\[ F_{lm} = F_{Fm}F_{-l-m}. \]  

Denoting the phases of the visibility and the bispectrum by \( \phi_n \) and \( \psi_{lm} \), respectively, replacing \( l \) by \( n - m \), and using Eq. (21) show that

\[ \phi_n = \phi_m + \phi_{n-m} - \psi_{n-m}. \]  

For \( m = 1 \), Eq. (22) may be used recursively to find all the \( \phi_n \) since \( \phi_0 \) and \( \phi_1 \) can be assigned arbitrarily. Furthermore, the procedure may be repeated with different values of \( m \) to obtain independent estimates of \( \phi_n \), which can be averaged, making the procedure much less sensitive to noise. The exponential factors \( \exp(i\phi) \) are usually averaged as a convenient means of accounting for the phases’ being modulo 2\( \pi \), so the averaged phases are given by

\[ \exp(i\phi_n) = (2/\pi) \sum_{m=1}^{n/2} \exp[i(\phi_m + \phi_{n-m} - \psi_{n-m})]. \]  

Note that the estimates of \( \phi_n \) are averaged before the value is used in the estimates of \( \phi_{n+1} \), minimizing the noise sensitivity. This recursive algorithm can be extended straightforwardly to the two-dimensional case,\(^{38}\) although its application is hampered by the large computational cost of computing the average bispectrum, which is a four-dimensional function. This technique can be viewed as a generalization of the Knox–Thompson procedure.

### 3. Phase Problem in Crystallography

X-ray crystallography is a technique for determining the structures of molecules. By structure is meant the position of each atom in the molecule. For an introduction to x-ray crystallography, the reader is referred to Ref. 38. Because of the short wavelength of x rays (~1 Å), molecules cannot be imaged directly. The diffraction pattern formed by the x rays scattered from a sample is therefore measured, and the data are processed numerically to determine the structure. Since a single molecule is extremely small, it is necessary to use a crystalline sample, which contains many molecules identically oriented and regularly spaced, to obtain measurable diffracted amplitudes. The scattering of the x rays by the sample is weak, so that the Born approximation applies and the scattered field is accurately approximated as the Fourier transform of the electron density \( f(x) \) in the crystal. Under normal circumstances, therefore, \( f(x) \geq 0 \). Only the intensity of the diffracted x rays can be measured, and so a phase problem arises. The differences between the crystallographic phase problem and that described in Section 2 arise from the crystalline nature of the sample.

A crystal can be described in terms of a unit cell that is repeated by multiple translations along three unit cell vectors in three-dimensional space (Fig. 1). If each unit cell is referenced to a point, the infinite set of points so formed is
called the lattice. The electron density in the crystal can therefore be written as

\[ f(x, y, z) = e(x, y, z) \otimes \sum_{m,n,p=-\infty}^{\infty} \delta(x - ma)\delta(y - nb)\delta(z - pc), \]  

where \( \otimes \) denotes convolution, \( \ell(x) \) is the lattice, the \( x_n \) are the lattice points, and \( e(x) \) denotes the electron density in a single unit cell (and zero elsewhere). There are different types of lattice, but, for ease of exposition, I will limit the discussion here to primitive orthorhombic lattices in which the unit cell vectors are mutually orthogonal and there is one molecule per unit cell. Equation (24) can then be written as

\[ f(x, y, z) = e(x, y, z) \otimes \sum_{m,n,p=-\infty}^{\infty} \delta(x - ma)\delta(y - nb)\delta(z - pc), \]  

where \( x = (x, y, z) \) and \( a, b, \) and \( c \) are the lengths of the unit cell vectors. The unit cell may have internal symmetries called space group symmetry, but, to avoid technical details not important in this paper, I will consider a unit cell with no symmetry. This is called space group P1. The unit cell contains \( N \), say, atoms, so that the electron density in the unit cell is given by

\[ e(x, y, z) = \sum_{n=1}^{N} e_n(x - x_n, y - y_n, z - z_n), \]  

where \( e_n(x, y, z) \) is the electron density, and \( (x_n, y_n, z_n) \) the position, of the \( n \)th atom. Since the electron density is concentrated around the atomic nuclei (Fig. 2), \( e(x) \) peaks at the atomic positions (this property is known as atomicity), and determining the electron-density map \( e(x) \) allows the atomic positions to be inferred. The type of atom can be determined from the amplitude of \( e(x) \), which is proportional to the atomic number. Note that it is strictly \( e(x) \), rather than \( f(x) \), that is to be reconstructed since \( f(x) \) is infinite in extent.

Since the x rays are diffracted by the whole crystal, the measured amplitude is \( |F(u)| = |F(u, v, w)| \), where, using Eqs. (1) and (25),

\[ F(u, v, w) = E(u, v, w) \sum_{h,k,l=-\infty}^{\infty} \delta(u - h/a)\delta(v - k/b)\delta(w - l/c), \]  

where \( E(u) \) is the Fourier transform of \( e(x) \), the diffraction pattern that would be recorded from a single unit cell. The measurements are therefore the Fourier transform of the electron density of a single unit cell sampled at the points \((h/a, k/b, l/c)\). Fourier space is called reciprocal space in crystallography, and the above points form the reciprocal lattice. Image space is referred to as real space. It is conventional to denote the samples of \( F(u, v, w) \), called the structure factors, by \( F_{hkl} \), i.e.,

\[ F_{hkl} = F(h/a, k/b, l/c) = E(h/a, k/b, l/c) = E(u)L(u), \]  

and \( L(u) \) denotes the reciprocal lattice. Note that \( \ell(x) \) and \( L(u) \) form a Fourier-transform pair. It is clear that

\[ F_{hkl} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z) \exp[i2\pi(hx/a + ky/b + lz/c)] dx dy dz \]  

or

\[ F_h = \int_{-\infty}^{\infty} f(x) \exp(i2\pi h \cdot x) dx, \]  

where the notation \( h = (h/a, k/b, l/c) \) is used to save space and \( V \) is the volume of the unit cell. From Eq. (26), therefore,

\[ F_{hkl} = \sum_{n=1}^{N} f_n \exp[i2\pi(hx_n/a + ky_n/b + lz_n/c)], \]  

where \( f_n \) is the atomic scattering factor of the \( n \)th atom [and is, implicitly, a function of \((h, k, l)\) when it is positioned at the origin. The electron density is given by

\[ f(x, y, z) = (1/V) \sum_{h,k,l=-\infty}^{\infty} F_{hkl} \exp[-i2\pi(hx/a + ky/b + lz/c)] \]  

or

\[ f(x) = (1/V) \sum_{h} F_h \exp(-i2\pi h \cdot x), \]  

although it is \( |F_h| \), not \( F_h \), that is measured.

An important point to note is that, because the Fourier amplitude is sampled, the arguments used in Section 2 to establish uniqueness of phase retrieval do not apply, since they make use of the Fourier transform and its amplitude as a continuous function in Fourier space. If the complex amplitude \( F_{hkl} \) could be measured, this would be of no significance because the samples are separated by the Nyquist spacing and are therefore sufficient to permit \( E(u, v, w) \) to be reconstructed continuously by the sampling theorem. However, the support of the autocorrelation in any direction is twice the support of \( e(x) \) in the same direction. The Nyquist spacing for \( |E(u)|^2 \) is therefore half that for \( E(u) \), so
that the continuous $|F(u)|^2$ cannot be reconstructed from the measured samples. The significance of this for crystallographic phase retrieval was first noted by Sayre. It follows that the autocorrelation of $e(x)$ cannot be reconstructed from the $|F_{hk}|^2$ either. What can be constructed is the Patterson map (also often called the Patterson) of $e(x)$, denoted by $P(x)$, which is the Fourier transform of the $|F_h|^2$, or the autocorrelation of $f(x)$:

$$P(x) = \frac{1}{V} \sum |F_h|^2 \exp(i2\pi h \cdot x).$$  \hspace{1cm} (32)

Using Eqs. (29) and (32) shows that

$$P(x, y, z) = \sum_{h,k,l=-\infty} A(x-ha, y-kb, z-\lambda c),$$  \hspace{1cm} (33)

where $A(x)$ is the autocorrelation of $e(x)$. Since $A(x)$ has an extent twice that of $e(x)$, the autocorrelations in Eq. (33) overlap, as shown in one dimension in Fig. 3. The sampling difficulty for $|F_{hk}|^2$ described above can therefore be interpreted as an aliasing problem in autocorrelation space. Since $e(x)$ is atomistic, so is $P(x)$, and peaks in $P(x)$ occur at positions that correspond to the interatomic vectors in the structure. In principle, $P(x)$ can be unrolled to obtain $e(x)$, but this is usually impossible unless the Patterson map is dominated by interatomic vectors among a few heavy atoms.

4. CLASSES OF CRYSTALLOGRAPHIC PHASE PROBLEM

The nature of phase retrieval in crystallography can be divided into a number of distinct classes. These result from implications of different sizes of molecules (the number of atoms), the quality (particularly the resolution) of the diffraction data, and the presence of disorder in the specimen. The three classes are

(1) Small molecules containing fewer than $\sim 150$ nonhydrogen atoms.

(2) Macromolecules with more than $\sim 500$ atoms. (Molecules of intermediate size are often solved by using a combination of different techniques.)

(3) Fiber diffraction in which the molecules are cylindrically disordered in the specimen.

The techniques used in these classes are described in the three following sections. The order in which they are presented is partly historical and also represents the theoretical power of the methods. The macromolecular class corresponds most closely to phase retrieval in general imaging, and I will therefore devote more attention to this class. Maximum-entropy methods have been explored in each of the three classes of crystallographic problem, and these are briefly mentioned at the end of each section.

5. SMALL-MOLECULE PHASE RETRIEVAL

The properties of small molecules that make this class of phase problem distinct are the small number of atoms in the unit cell and the availability of high-resolution diffraction data. The latter property means that the diffracted amplitudes $|F_{hk}|^2$ are measured out to distances from the origin in reciprocal space that correspond to at least the minimum distance between atoms in the structure ($\sim 1.5$ Å). The importance of this property is that the atomicity of the structure is then embodied in the data. The importance of the first property is described below. Almost all small-molecule structures are solved now by using direct methods, but it is instructive to consider first the heavy-atom method that was used before the development of direct methods and is still useful in some circumstances.

Heavy-Atom Method

The heavy-atom method\cite{38} is based on solving the structure of a molecule that contains a heavy atom (either naturally or one that is artificially introduced), i.e., an atom that has a much larger atomic scattering factor that the other atoms in the molecule. This has two effects: The phases of many of the structure factors are dominated by the phases that are due to the heavy atom, and the Patterson function is dominated by peaks corresponding to interatomic vectors between the heavy atoms. Inspection of the Patterson map allows the positions of the heavy atoms in the unit cell to be determined. The structure factor phases that are due to the heavy atom(s) alone are then calculated and combined with the measured amplitudes to calculate an approximate electron-density map, using Eq. (31). This usually allows additional atoms to be located and thence more-accurate phases to be calculated. The process is repeated with atoms being added at each iteration until all have been located.

Direct Methods

Most small-molecule structures are solved now by using direct methods. These methods are so called because they utilize the structure amplitudes alone to calculate an initial set of phases. Direct methods are a powerful collection of techniques, and their importance is attested to by the award of a Nobel Prize to J. Karle and H. Hauptman, two of the pioneers of the method, in 1985. The reader is referred to Refs. 40 and 41 for detailed reviews of direct methods. The development of direct methods began in 1948 when Harker and Kasper\cite{12} used the Cauchy inequality to derive inequalities that, in some cases, could give phase information. Shortly afterward, Karle and Hauptman\cite{42} showed that non-negativity of the electron density requires that any determinant of the following form satisfy

$$\begin{vmatrix} F_0 & F_{-h_1} & F_{-h_2} & \cdots & F_{-h_n} \\ F_{h_1} & F_0 & & & \\ \vdots & \vdots & \ddots & \vdots \\ F_{h_n} & F_{h_n+h_1} & \cdots & F_0 \end{vmatrix} \geq 0. \hspace{1cm} (34)$$

In 1962 Sayre\cite{44} (also one of the founders of direct meth-
odds) showed that for a structure containing identical resolvable atoms,
\[ F_h = \theta_h \sum_k F_k F_{h-k}, \]  
(35)

where \( \theta_h \) can be determined from the atomic scattering factors and the sum is over all values of \( k \). Sayre’s equation represents the first time that the atomicity property was utilized and is beautiful in its simplicity. Despite the requirement of identical atoms, Eq. (35) is often approximately satisfied for structures containing different types of atom. Sayre’s equation is used in direct methods to assess the reliability of phase relationships.

Expanding the third-order determinant [relation (34)] shows that
\[ F_h^2 \leq (1 + F_{3h})/2, \]
(36)

which are known as sigma-one relationships. They sometimes, but not often, allow the phase of \( F_{3h} \) to be determined from the magnitudes only. Utilizing the third-order determinant again and considering structure factors of unusually large magnitude show that
\[ F_h \approx F_k F_{h-k} / F_0, \]
(37)

which implies that
\[ \phi_h = \phi_k + \phi_{h-k}, \]
(38)

which are called triplet, or sigma-two, relationships. They state that, for large structure factor magnitudes, the phase of one reflection may be defined, approximately, in terms of the phases of two other reflections. In essence, the (exact) Sayre equation [Eq. (35)] for large \( F_h \) is dominated by the tripies [relations (37) and (38)] where \( F_k \) and \( F_{h-k} \) are both relatively large.

Atomicity is introduced by treating the atomic positions as (uniformly distributed) random variables. This has turned out to be a powerful technique and is the basis of modern direct methods. The first developments in direct methods applied to centrosymmetric structures [for which \( f(x) = f(-x) \)], but these will be skipped here, as they have little relevance to general imaging. In the probabilistic treatment, it is convenient to define the normalized structure factors \( E_h \) by
\[ E_h = F_h \left( \sum_{j=1}^N |f_j|^2 \right)^{1/2}. \]
(39)

This is called sharpening, as it makes the atomic electron densities appear more like delta functions. The triplet relationship [relation (38)] has found wide application beyond the expected range (large structure factor magnitudes) of its validity because of the probabilistic characteristics of the inequalities. Calculating joint probability distributions of structure factors shows that the probability-density function associated with relation (38) is given by
\[ P(\Phi_3) = \exp(B \cos \Phi_3) / [2 \pi I_0(B)], \]
(40)

where
\[ \Phi_3 = \phi_h - \phi_k - \phi_{h-k} \approx 0, \]
(41)

\( I_0 \) is the zero-order modified Bessel function of the first kind, and

The variance of \( \Phi_3 \) decreases with increasing \( B \), so that relation (38) is more accurate for fewer atoms in the unit cell and larger values of the normalized structure amplitudes (Fig. 4). In the later stages of the phasing process a number of relationships of the form of relation (38) usually relate to \( \phi_h \) and the expression
\[ \tan \phi_h = \sum_k \frac{|E_k E_{h-k}| \sin(\phi_k + \phi_{h-k})}{\sum_k |E_k E_{h-k}| \cos(\phi_k + \phi_{h-k})}, \]
(43)

(which can be derived straightforwardly from Sayre’s equation), where \( k \) is restricted to values such that \( |E_k| \) and \( |E_{h-k}| \) are large, has a higher probability of being correct than any of the individual terms in relation (38). The variance of \( \phi_h \) derived from Eq. (43) can be calculated and used as a figure of merit. Equation (43) is called the tangent formula, and it plays an important role in direct methods.

With the results described above, phase determination proceeds as follows: The structure amplitudes are converted to normalized structure amplitudes, and sigma-one relationships are applied if possible. A set of reflections with the largest \( |E_k| \) values is chosen, the possible triples are listed, and \( B \) is calculated for each triple as a first evaluation of its efficiency. Reflections are chosen for fixing the origin and the enantiomorph \( f(x) \) or \( f(-x) \), and these form the starting phase set. Since origin fixing and sigma-one relationships may not be sufficient to start phase determination, symbolic phases are usually associated with the high modulus reflections, permitting relationships between phases to be carried through before the actual values are determined. This is known as symbolic addition and was important in making direct methods practical. Phases are assigned to other reflections, by using first the triplet relationships and then the tangent formula. Some phases may remain as symbols, and some may be assigned numerical values. Actual values are assigned to symbols as early as possible, and if \( p \) symbols are each assigned phase values \( \pi/4, 3\pi/4, 5\pi/4, \) and \( 7\pi/4 \), then in one of the \( 4^p \) solutions each symbolic phase has unit probability of being within \( 45^\circ \) of the correct value. The phases are then refined and extended (to other reflections) by (repeated) use of the tangent formula. A set of reflections with phases that have the highest figure of merit is used to calculate an electron-density map. Enough atoms are usually located in the map to begin refinement (adjust-
ment of the atomic positions to optimize the agreement between the measured amplitudes and those calculated from a model of the structure. Usually at least a fragment of the structure is visible, and this may be used to calculate a set of structure factors. The phases of the large structure factors are taken as correct and tangent formula refinement and extension repeated. Repeated application usually results in a recognizable electron-density map. Difference Fourier synthesis, in which the coefficients \( F_k \) in Eq. (31) are replaced by the differences between the measured and calculated (from the estimated structure) structure amplitudes together with the calculated phases, is useful. Maps calculated in this way show (approximately) the differences between the current model and the actual structure—i.e., the errors in the model. These methods are integrated into a program package called MULTAN (the multiple-tangent-formula method). A weighted tangent formula is often used in which the terms in Eq. (43) are weighted, depending on the estimated reliability of the phases they contain. An interesting random approach has proved powerful and has been incorporated into a program called RANTAN (random MULTAN). In this method, a large starting set is taken, and, for each trial, random phases are assigned to the reflections and refined by using the weighted-tangent formula until good figures of merit are obtained. This often produces a high-quality electron-density map faster than MULTAN can.

Many refinements and new techniques have been incorporated into direct methods or are currently under investigation, and I will not attempt to cover them here. They include the use of higher-order invariants (such as quartets and quintets) and use of the neighborhood principle and further applications of the Sayre equation and Karle-Hauptman determinants. Direct methods are shown by Bricogne to bear a close relationship to a maximum-entropy formulation. He shows that a scheme that involves updating a (nonuniform) prior distribution of atoms as phase determination proceeds can be formulated as a maximum-entropy problem. This is a potentially powerful approach to extending the scope of direct methods (to larger molecules), although more study is required to assess the potential. He also shows that this formalism is related to the "maximum determinant method," which is a probabilistic result involving the Karle-Hauptman determinants [Eq. (34)].

6. MACROMOLECULAR PHASE RETRIEVAL

Macromolecular crystallography deals with determining the structures of molecules that contain more than ~500 atoms. The term protein crystallography is synonymous with macromolecular crystallography since most of the structures determined have been proteins, although the methods have also been applied to nucleic acids and to protein–nucleic acid complexes including a number of viruses. Phase retrieval in macromolecular crystallography is quite different from that for small molecules. Direct methods are not effective for two reasons: First, and more significantly, the large number of atoms results in the variances of the phase probabilities being so large that the phase-predicting power is negligible. Second, diffraction data are usually measurable only to moderate resolution (~2–3 Å), so that the atomicity property, on which direct methods are based, is not reflected in the data. All procedures used in macromolecular crystallography therefore utilize some ancilliary information that allows approximate initial phases to be calculated. This information is obtained with a variety of experimental techniques or by using a known similar structure. The initial phases are then refined and extended by using a number of algorithms. I briefly describe here the experimental procedures used to determine initial phases and the phase-refinement and -extension algorithms.

Isomorphous Replacement

The method of isomorphous replacement was first used in the 1950's and is still central to the analysis of macromolecules as a means of obtaining initial phase information. An isomorphous derivative of the crystal is prepared in which a heavy atom is attached to the macromolecule. The heavy atom introduces measurable changes in the diffraction pattern, and diffraction data are collected from both the native crystal and the derivative. If \( F_h \), \( F_h^D \), and \( F_h^H \) are the structure factors for the native, the derivative, and the heavy atom, respectively, then

\[
F_h^D = F_h + F_h^H, \tag{44}
\]

and putting this in the form of amplitudes and phases shows that (Fig. 5)

\[
\phi_h = \phi_h^H + \cos^{-1}[(|F_h|^2 - |F_h^H|^2)/(2|F_h||F_h^H|)]. \tag{45}
\]

The phases \( \phi_h \) can therefore be determined if the structure factors of the heavy atom are known. The position of the heavy atom is determined by using an isomorphous difference Patterson map that is calculated by using the coefficients 

\[
P(x) = (1/u) \sum_h (|F_h^D| - |F_h|^2) \exp(i2\pi h \cdot x). \tag{46}
\]

The coefficients should ideally, of course, be \( |F_h|^2 \), but \( |F_h^D| - |F_h|^2 \) is a good estimate of \( |F_h|^2 \) if \( F_h^D \) and \( F_h \) have similar phases. The phases tend to be similar when the difference \( |F_h^D| - |F_h|^2 \) is large, so that the sum in Eq. (46) is restricted to those values only, while retaining a reasonable number of terms. The Patterson map is dominated by peaks between the heavy atoms and can thus be used to determine their positions, allowing \( F_h^H \) (magnitudes and phase) to be calculated. All the quantities on the right-hand side of Eq. (45) are then known, allowing \( \phi_h \) to be calculated. Because of the arccosine in Eq. (45), there is a twofold ambiguity in the value of \( \phi_h \) however. This can be resolved by using two different isomorphous derivatives and choosing the phase solutions that coincide.

In practice, to help to ameliorate

Fig. 5. Relationships between the structure factors used to derive Eq. (45) for isomorphous replacement.
the effects of noise a number of different isomorphous derivatives are used, and this is called multiple isomorphous replacement (MIR). The approximate phases so obtained are combined in an optimum manner (as described below) and used to calculate an electron-density map. The phases are often reasonably accurate to ~3-Å resolution, and the resulting map is often interpretable, i.e., the chemical structure of the macromolecule (which is usually known) can be rationalized with the electron-density distribution. The known chemical structure is then built into the map (using computer graphics and special-purpose software) to obtain an approximate three-dimensional model. The model is refined by adjusting it to optimize the agreement between the measured and calculated amplitudes (as well as optimizing the stereochemistry) by using, for example, conjugate gradient algorithms.\textsuperscript{57} Difference Fourier synthesis is also used during refinement to detect errors in the model.

If it is possible experimentally to obtain only a single isomorphous derivative, then it may be possible to calculate a useful (but noisy) electron-density map by using coefficients $|F_h| \exp(i\phi_h') + |F_h| \exp(i\phi_h'')$ in Eq. (31), where $\phi_h'$ and $\phi_h''$ are the two phase solutions obtained from Eq. (45). A map calculated in this way usually has some correct features that are due to the correct phases and a noisy background that is due to the incorrect phases. Such a map may be interpretable or may be good enough to begin density modification (described below) with. This is called single isomorphous replacement (SIR), the obvious advantage being that, if it is successful, the amount of experimental work will be substantially reduced.

In practice, errors are present owing to inaccuracies in intensity measurements, lack of isomorphism, and imprecise heavy-atom positions. An important part of the procedure is to take account of the errors to obtain the best phase estimate—that which gives the least mean-square error in the electron-density map. Assuming normally distributed errors, the probability that a phase angle $\phi$ is correct is given by

$$P(\phi) = \exp\left\{-\frac{\epsilon(\phi)^2}{2E^2}\right\},$$

(47)

where the subscript $h$ has been dropped, $\epsilon$ is the lack of closure of the phase triangle (Fig. 6) formed by $F, F_0^D,$ and $F_0^{\text{II}}$, and $E$ is an estimate of the total error.\textsuperscript{58} The lack of closure can be estimated since $|F|$ and $|F_0^D|$ have been measured, and $|F_0^{\text{II}}|$ and the angle $(\phi - \phi^{\text{II}})$ between $F$ and $F_0^{\text{II}}$ have been estimated by the isomorphous replacement procedure. When several derivatives are used, the total probability is given by

$$P(\phi) = \exp\left\{-\sum_j \left[\epsilon_j(\phi)^2/2E_j^2\right]\right\}.$$  

(48)

It can be shown\textsuperscript{58} that the electron-density map with the least mean-square error is obtained by using the coefficients $m|F_h| \exp(i\phi_h)$ in Eq. (31), where the best phase $\phi_{\text{best}}$ and the figure of merit $m$ are given by

$$\tan(\phi_{\text{best}}) = \int_0^{2\pi} P(\phi) \sin \phi d\phi / \int_0^{2\pi} P(\phi) \cos \phi d\phi$$

(49)

and

$$m = \int_0^{2\pi} P(\phi) \cos(\phi_{\text{best}} - \phi) d\phi / \int_0^{2\pi} P(\phi) d\phi.$$  

(50)

Note that $m = 1$ corresponds to zero phase error and $m = 0$ corresponds to $\pi/2$ phase error (random phases). A slightly different treatment of phase probabilities given by Hendrickson and Lattman\textsuperscript{59} has a number of practical advantages.

**Anomalous Scattering**

Another experimental method for obtaining approximate phases is anomalous scattering. If the wavelength of the x rays used is close to an absorption edge of an atom, the scattered x rays will suffer dispersion and absorption. The effect of this is that the scattering factor for the atom changes, becomes complex, and is given by

$$f = f_o + \Delta f' + i\Delta f''$$

(51)

where $f_o$ is the ordinary scattering factor and the components $\Delta f'$ (due to dispersion) and $\Delta f''$ (due to absorption) depend on the type of atom and the x-ray wavelength. In practice, the effects of anomalous scattering are significant only for a few heavy atoms in the structure. Two classes of technique are in use: one in which absorption differences at a single wavelength are used to complement SIR and another that uses multiple wavelength measurements on the native molecule. For simplicity, consider the case in which there is only one type of anomalous scatterer; let the real and the imaginary parts of the atomic scattering factors be $f' = f_o + \Delta f'$ and $f'' = \Delta f''$, respectively, and the structure factors due to $f'$ and $f''$ be $|F_h| \exp(i\phi_h')$ and $|F_h| \exp(i\phi_h'')$, respectively.

Since the scattering factor is complex, the usual condition $|F_h| = |F_{-h}|$ is not satisfied, and the difference $|F_h| - |F_{-h}|$ is known as the Bijvoet difference. An analysis similar to that for isomorphous replacement shows that, to a good approximation,\textsuperscript{60}

$$|F_h| - |F_{-h}| = 2|F_h| \sin(\phi_h'' - \phi_h'^{-}),$$

(52)

where the superscript $A$ indicates the contribution from the anomalous scatterers (Fig. 7). The anomalous scatterer positions can be determined by using an anomalous difference Patterson map, which is similar to the isomorphous difference Patterson map. All the quantities in relation (52), except $\phi_h'$, can be measured or calculated. In the single-wavelength method, solution of relation (52) gives two values for $\phi_h$, which are combined with SIR phases to obtain a best phase as described above, although the effects of errors have to be treated carefully.\textsuperscript{61}

In the multiple wavelength method, called multiple anomalous dispersion (MAD), measurements are made from a single specimen containing an anomalous scatterer(s) at a number (usually 3–5) of x-ray wavelengths. For each pair of wavelengths it is possible to set up equations of the form\textsuperscript{60}
Density modification is a general term for methods of phase determination that incorporate known constraints on the electron density. These techniques usually make use of noncrystallographic symmetry and solvent leveling. Unfortunately, the term molecular replacement (above) is also used to describe density modification. (One must be aware of this when reading the literature.) Many proteins and viruses have multiple-subunit structures in which the molecule is made up of identical subunits with positions related by symmetry operations. If these symmetry operations coincide with the crystal (space group) symmetries, then no advantage is gained. However, sometimes the symmetry operations do not coincide with the space group symmetries, and these symmetries are called noncrystallographic. Noncrystallographic symmetry is local in that it applies only within a restricted volume surrounding the molecule, whereas crystallographic symmetry applies to the whole crystal. Rossmann and Blow recognized that the presence of noncrystallographic symmetry may provide phase information. A good review of these techniques is Ref. 62.

The presence of noncrystallographic symmetry may be used to derive equations relating the structure factors. Consider a molecule made of two identical subunits with positions and orientations related by noncrystallographic symmetry so that

\[ f(x) = f(Cx + d), \quad x \in U, \]

where \( C \) and \( d \) are a rotation matrix and a translation vector, respectively, that relate the position and orientation of the second subunit to the first and \( U \) is the molecular envelope that encloses the local region in which the noncrystallographic symmetry applies. Using Eqs. (55) and (31) shows that

\[ f(x) = (1/\delta) \sum_{\mathbf{h}} F_{\mathbf{h}} \exp[-i2\pi \mathbf{h} \cdot (C\mathbf{x} + \mathbf{d})], \]

and substituting for \( f(x) \) from Eq. (56) into Eq. (29) gives

\[ F_{\mathbf{p}} = \sum_{\mathbf{h}} H_{\mathbf{hp}} F_{\mathbf{h}}, \]

where

\[ H_{\mathbf{hp}} = \exp(-i2\pi \mathbf{h} \cdot \mathbf{d}) \int_{U} \exp[i2\pi (\mathbf{p} - \mathbf{hC}) \cdot \mathbf{x}] d\mathbf{x}. \]

The matrix elements \( H_{\mathbf{hp}} \) depend only on the rotation and translation parameters and on the molecular envelope. In practice, the crystallographic symmetry must also be considered, which gives more-complicated equations of the same form as Eqs. (57) and (58). For \( N \)-fold noncrystallographic symmetry, \((N - 1)\) sets of equations of the form of Eq. (57) are obtained. The system of equations, for different \( \mathbf{p} \), represented by Eq. (57) are called the molecular replacement equations and can, in principle, be solved for the phases of \( F_{\mathbf{h}} \). However, solution of the equations as a method of ab initio phase determination has been largely unsuccessful owing to the difficulty of solution and the large computational cost involved. A practical solution is real space averaging, described below. In order to make use of noncrystallographic symmetry, one must determine the rotation and translation parameters. Determining these parameters is referred to as the rotation and translation problems, respectively, and I will describe these next.

The rotation problem was first addressed by Rossmann and Blow, who defined the rotation function \( R(\theta_1, \theta_2, \theta_3) \) by

\[ R(\theta_1, \theta_2, \theta_3) = \int_{U} P(x) P(y) dx, \]

where \( y \) represents \( x \) rotated through the Eulerian angles \( (\theta_1, \theta_2, \theta_3) \) and \( P(x) \) is the Patterson function. Using Eq. (32) allows Eq. (59) to be written as

\[ R(\theta_1, \theta_2, \theta_3) = \sum_{\mathbf{p}} |F_{\mathbf{p}}|^2 \sum_{\mathbf{n}} |F_{\mathbf{n}}|^2 G_{\mathbf{pn}}, \]

where

\[ |F_{\mathbf{p}}|^2 = |F_{\mathbf{h}}|^2 \sum_{\mathbf{u}} H_{\mathbf{hp}}^* H_{\mathbf{iu}}. \]
and the matrix \( C \) represents the rotation defined by \((\theta_1, \theta_2, \theta_3)\). The rotation function can therefore be calculated from the measured amplitudes by using Eqs. (60) and (61). Computation of the rotation function is considerably accelerated by expanding the Patterson functions in spherical harmonics. The Patterson function rotated relative to itself gives a maximum overlap of the interatomic vectors when one subunit has the same orientation as another subunit. Hence the rotation function has peaks in directions \((\theta_1, \theta_2, \theta_3)\) corresponding to the relative orientations of subunits and is used to determine the rotation components of the noncrystallographic symmetry.

Once the rotation part of the noncrystallographic symmetry is determined, the translation part needs to be found. The translation problem is less tractable than the rotation problem and has been discussed by a number of authors. In the special case that the noncrystallographic rotation is twofold, the translation vector can be determined by using a translation function similar in form to the rotation function except that the Patterson functions are translated rather than rotated. If a related structure is known (as in applications of molecular replacement), then various translation functions may be used. Alternatively, possible molecular positions may be assessed by examining their effects on the agreement between calculated and measured structure amplitudes and/or by packing considerations (usually some molecular positions may be excluded because they would involve interpenetration of adjacent molecules). If approximate, low-resolution phases are available (from isomorphous replacement, for example), inspection of a calculated electron-density map may be sufficient to position the molecule. The translation problem can often be solved by using a combination of these techniques.

Although solution of the molecular replacement equations has not been practical, so called (electron) density modification (also called real-space molecular replacement) has proved to be a powerful technique and has been an essential part of some high-resolution protein and, particularly, virus structure determinations. The procedure involves iterating between real space and Fourier space, enforcing structural information (noncrystallographic symmetry, for example) in real space, and ensuring consistency with the measured amplitudes in reciprocal space. Density modification seeks to refine and extend the phases rather than to attempt \(\text{ab initio}\) phase determination.

The reader probably has recognized that this procedure has the flavor of the (Fienup) iterative phase-retrieval algorithms described in Section 2. The constraints typically available in real space are noncrystallographic symmetry (equivalence of subunits), the molecular envelope (a form of support constraint), and positivity of the electron density. Reciprocal space constraints are the measured structure amplitudes. An outline of the density modification procedure is as follows:

1. Calculate approximate low-resolution phases, using MIR, SIR, anomalous scattering, from a related structure, or some combination of these.
2. Determine the noncrystallographic symmetry parameters (using the rotation and translation functions) and an approximate envelope from inspection of a low-resolution electron-density map derived from the phases from step 1.
3. Calculate an electron-density map, using the current phases and weighted measured amplitudes.
4. Enforce real-space constraints by
   a. Setting the electron density in each subunit to the average value over all the subunits.
   b. Setting the electron density outside the envelope to zero or to the average solvent value (this is called solvent leveling).
   c. Setting any negative electron density values to zero or to the average solvent value.
5. Calculate a new set of phases, and synthesize them with the isomorphous replacement phases to obtain a new combined best phase.
6. Return to step 3.

This procedure is iterated until there is little change in the phases. The progress of refinement is monitored by examining the agreement between the calculated and measured structure amplitudes; failure of an agreement index to attain a reasonable value indicates that the phases are not improving sufficiently. The success of phase refinement depends on the accuracy of the structure amplitude measurements and the starting phases, on the degree of noncrystallographic symmetry, and on the molecular envelope fractional volume \((U/V)\). If a molecule has noncrystallographic symmetry, then the phases obtained by isomorphous replacement do not need to be so accurate, since they can be refined to more accurate values with this procedure. This eases the amount of experimental work required and has been used particularly successfully to refine SIR phases. It can be shown that this procedure is equivalent to iterative solution of the molecular replacement equations.

Once the phases have converged, higher-resolution reflections can be phased by phase extension, which is most effectively performed by adding reflections in a spherical shell in reciprocal space of thickness equal to a few reciprocal lattice points. The current amplitudes and calculated phases are then used to repeat the above procedure from step 3, to phase the shell of reflections added. Once this procedure has converged, another shell of reflections is added. The extension is continued until poor agreement between the calculated and measured amplitudes indicates that the accuracy of the high-resolution data is deteriorating. In practice, there are always regions of reciprocal space where the structure amplitudes cannot be measured, and it is important to use the current calculated amplitudes (rather than excluding the reflections) to stabilize and accelerate convergence. Phase determination is complete if the phases can be extended to high enough resolution for the electron-density map to be interpretable. A model is then built into the electron-density map and the model refined as described above in the subsection on isomorphous replacement.

Maximum entropy has been explored in macromolecular phase retrieval as a means of improving the quality and resolution of electron-density maps and of incorporating a variety of different types of data. A few applications have been reported in which the phases have been determined.
and the resolution extended. Although there appear to be improvements in these maps, more study is needed to assess the value of these methods.

7. PHASE RETRIEVAL IN FIBER DIFFRACTION

Certain molecules, such as polymers and macromolecules made up of units that aggregate with helix symmetry, do not form regular three-dimensional crystals as described in Section 3 and are therefore not amenable to conventional crystallographic analysis. However, these molecules can often be prepared as fibers in which their long axes are approximately parallel, but the molecules are subject to random rotation about their long axes, and structure determination is possible by using x-ray diffraction from such specimens. The techniques used are called fiber diffraction and can be considered a variation of traditional crystallography. Fiber diffraction was used in determining the double-helix structure of DNA in the 1950's.

Since the molecules are randomly rotated, it is convenient to use cylindrical coordinates \((r, \phi, z)\) and \((R, \psi, Z)\) in real and reciprocal space, respectively. Fibrous molecules always display helix symmetry in which a subunit is repeated indefinitely by helix symmetry operations. The helix symmetry is denoted by \(\omega_n\), which means that there are \(u\) subunits in \(v\) turns of the helix. The \(z\) axis is defined as coinciding with the helix axis, and the helix pitch is \(P\), so that the repeat distance \(c = vP\). In cylindrical coordinates, the Fourier transform of a single molecule is given by

\[
F(R, \psi) = (1/c) \sum_l \sum_n f_j \sum_n f_j (2\pi R r_j) \exp[i(n(\psi + \pi/2) - n\phi_j + 2\pi l z/c)],
\]

(62)

which is restricted to layer planes \(z = l/c\), where \(l\) is an integer, \(J_n\) is the \(n\)th-order Bessel function of the first kind, and \((r_j, \phi_j, z_j)\) are the coordinates of the \(j\)th atom in one subunit. The sum is over all integers \(n\), but the helix symmetry results in all terms other than those satisfying the helix selection rule \(l = um + vn\), where \(m\) is any integer, being equal to zero. The random rotation of the molecules results in cylindrical averaging of the diffraction pattern so that the quantity measured is \(I_l(R)\), given by

\[
I_l(R) = (1/2\pi) \int_0^{2\pi} |F(R, \psi)|^2 d\psi,
\]

(63)

where \(l\) indexes the layer planes. Making use of Eq. (62) shows that Eq. (63) can be put in the form

\[
I_l(R) = \sum_n |G_n(R)|^2,
\]

(64)

where

\[
G_n(R) = \sum f_j \sum_n f_j (2\pi R r_j) \exp[i(2\pi l z/c)].
\]

(65)

The corresponding inverse transform for reconstructing the electron density \(f(r, \phi, z)\) is given by

\[
f(r, \phi, z) = \frac{1}{c} \sum_l \sum_n g_n(r) \exp[i(n\phi - 2\pi l z/c)],
\]

(66)

where

\[
g_n(r) = \frac{1}{2\pi} \int_0^{2\pi} g_n(R) J_n(2\pi R r) 2\pi R dR,
\]

(67)

so that it can be determined from the \(G_n(R)\). Hence the phase problem is much more difficult than in conventional crystallography since it involves separating out the individual \(|G_n(R)|^2\) terms in Eq. (64) from the measurements \(I_l(R)\) as well as determining the phase of each \(G_n(R)\).

It is important to note that, although the index \(n\) in Eq. (64) ranges over an infinite set, the behavior of the Bessel functions near the origin ensures that only a finite number, \(N\), say, are significant in practice. \(N\) depends on \(u\) and on the maximum values of \(r_j\) (i.e., the molecular radius) and \(R\). The implications of this for the phase problem have been discussed by Makowski, and, from the discussion above, it means that \(2N\) values (the real and the imaginary parts of each \(G_n(R)\)) need to be determined for each measurement of \(I_l(R)\). Typical values for \(N\) are between 3 and 8.

The method of isomorphous replacement described in Section 6 has been modified for application to fiber diffraction. In principle, \(2N\) derivatives are needed to separate and phase the \(G_n(R)\), although it can be difficult to obtain this many derivatives in practice. Let \(G_n^D\) and \(G_n^H\) be the \(G_n\) for the derivative and the heavy atom alone, respectively, so that

\[
G_n^D = G_n + G_n^H = (A_n + iB_n) + (A_n^H + iB_n^H),
\]

(68)

where the \(A_n\) and \(B_n\) are real and the dependence on \(l\) and \(R\) has been dropped to simplify the notation. The measured intensities \(I\) and \(I^D\) are therefore given by

\[
I = \sum_n (A_n^2 + B_n^2)
\]

(69)

and

\[
I^D = \sum_n [(A_n + A_n^H)^2 + (B_n + B_n^H)^2].
\]

(70)

The positions of the heavy atoms cannot be determined by Patterson methods in fiber diffraction, but they can be located by using the diffraction data for small \(R\) where only one \(G_n\) term contributes, so that \(A_n^D\) and \(B_n^D\) can be calculated. Since \(I\) and \(I^D\) are measured, Eqs. (69) and (70) form a system of quadratic equations in the \(2N\) unknowns \(A_n\) and \(B_n\). This system can be solved by a minimization procedure that finds, in addition to the global minimum, the local minima as well. An advantage of the diffraction data's being available continuously in \(R\), rather than only at the reciprocal lattice points as in conventional crystallography, is that a reliable most-probable phase, rather than the best phase described in Section 6, can be determined. This is done by using the continuity of \(G_n(R)\) as a function of \(R\); i.e., since the diffracting particle has a finite diameter, \(G_n(R)\) is band limited so that its rate of change with \(R\) is restricted by a maximum value. The most probable phase of \(G_n(R)\) is traced in the complex plane as a function of \(R\). If the change
from one sample of $R$ to the next exceeds the maximum value allowed, then the local minima described above are searched to find one that does not exceed the maximum change. This allows some of the false solutions, which are due to noise, to be detected and corrected. If fewer than $2N$ derivatives are available, but there are small deviations from the ideal helix symmetry, a phenomenon called layer-line splitting can be used to help to determine the ratios of the $|G_n(R)|^2$ terms in Eq. (64).

Once the $G_n(R)$ terms have been separated and phased, an electron-density map can be calculated by using Eqs. (66) and (67). The phases can then be refined with solvent leveling as described in Section 6. If the map is interpretable, an atomic model can be fitted and refined. Other techniques such as the use of a limited number of derivatives and difference Fourier synthesis have been adapted for use in fiber diffraction.

Low-resolution fiber diffraction structures have also been determined without any explicit starting phases (from isomorphous replacement) by using a density modification procedure in which solvent leveling and enhancement of large-scale features known to exist (from other independent information) in the structure are applied in alternate cycles. Solvent leveling is more powerful in fiber diffraction than in conventional crystallography, as I describe in Section 8. If such a procedure converges to the same structure for a wide variety of starting models, one can be reasonably confident that the final structure is correct. Without more-explicit phase information, this procedure is useful only at low resolution.

The types of specimen described above are called noncrystalline because there is no crystallinity in a plane normal to the helix axis. Many polymers form polycrystalline fibers in which the molecules organize laterally into microcrystallites, each of which is a small three-dimensional crystal, the crystallites being randomly rotated. In this case, therefore, the diffraction pattern is the cylindrically averaged sampled transform, so that, when the measurements are denoted by $H_i$,

$$H_i = \sum |F_{hkll}|^2,$$

where the sum is over the reciprocal lattice points $(h, k, l)$ that have the same cylindrical radius $R$. Many polymers with smaller subunits crystallize in this manner, and, with the addition of stereochemical information, the number of degrees of freedom in a structural model can often be reduced substantially. This allows various atomic models to be constructed and refined against the diffraction data and stereochemical constraints by using, for example, a least-squares procedure. The incorrect optimized models can often be detected by using various figures of merit, indicating which is the correct solution. In this method, the initial atomic model effectively provides approximate phases. This technique has, in fact, been used to solve most polymer structures, as it avoids the quite difficult and time-consuming preparation and analysis of derivatives. It is not effective, however, with larger structures (that have many degrees of freedom), for which phase estimates must be obtained with the methods described above before an atomic model can be constructed.

The maximum-entropy method has been applied to one fiber diffraction problem to obtain the most nearly uniform map of a helical virus consistent with model information and medium resolution diffraction data. It is not clear, however, whether this analysis produced a reliable map.

8. COMPARISON OF OPTICAL AND CRYSTALLOGRAPHIC PHASE RETRIEVAL

Fundamental Differences

There are, of course, many obvious common aspects of phase retrieval in crystallography and optics simply because one is dealing with a phase problem in both cases. I will summarize here the fundamental differences between the two cases and their implications.

The main differences between these two applications of phase retrieval arise because of the undersampling of the continuous intensity in crystallography compared with optics. The most important effect of this is on uniqueness. The uniqueness results in general imaging are based on the availability of the continuous transform. Note that this does not mean that the intensity must be measured continuously but only that the measured samples are sufficient, in principle, to reconstruct it. Since this is not the case in crystallography, phase retrieval is nonunique here in general. The success of direct methods in crystallography, however, provides strong evidence for uniqueness when atomicity is present and the number of atoms is small. Atomicity therefore appears to be a strong "a priori" constraint and to be more effective than positivity. Apart from arguments based on overdetermination (the ratio of the number of measured structure amplitudes to the number of unknown atomic positions is large), however, there is no definitive proof of uniqueness.

The macromolecular phase problem is nonunique, since atomicity is not in effect, unless other constraints, such as noncrystallographic symmetry, are present. In this case, therefore, one needs additional experimental information to estimate phases.

There are also implications of the crystallographic case's being three, rather than two, dimensional. In the optical (continuous) case, increasing the dimensionality increases the number of constraints on the intensity. The result is that the one-dimensional problem is nonunique (underdetermined), the two-dimensional problem is unique (but not overdetermined), and the three-dimensional case is overdetermined by a factor of $2^5$. In the crystallographic case, the problem is underdetermined by a factor of 2 from the point of view that each magnitude must be supplemented by its phase for the electron density to be reconstructed. However, if the additional information is in the form of additional amplitudes (from structural redundancy, for example), then it is more underdetermined. The sample spacing is twice the Nyquist spacing for each dimension, which leads to an eightfold reduction in the sampling density (in three dimensions). Considering the twofold overdetermination in the continuous case, the crystallographic problem is therefore underdetermined by a factor of 4. The effects of other structural constraints in crystallography have been discussed. These arguments are based on analyticity of the transform and do not take account of positivity, which would
be expected to increase overdeterminacy or uniqueness. Unfortunately, although conditions for image positivity in Fourier space can be formulated [as positivity of the Karle-Hampartner determinants [relation (34)], for example], it is difficult to formalize the effect of positivity on uniqueness. Positivity is not, in principle, necessary for uniqueness, but it is important in practice since successful reconstruction of general complex images requires a tight support.

Support constraints in crystallography (solvent leveling) are not so effective as in general imaging owing to the periodic image or, equivalently, to the undersampling of the continuous transform. A measure of the effectiveness of a support constraint in general imaging is the ratio of the area of the autocorrelation support to the area of the image support, which is equal to 4. In crystallography, an equivalent measure is the ratio of the unit cell volume to the envelope volume \((V/U)\), which is typically between 1 and 2 and thence not so effective.\textsuperscript{55,56}

The most powerful constraint available in macromolecule crystallography, if it is present, is noncrystallographic symmetry. The presence of noncrystallographic symmetry effectively reduces the undersampling of the continuous transform in reciprocal space.\textsuperscript{88} This is because noncrystallographic symmetry produces symmetry in reciprocal space that maps reciprocal lattice points to nonreciprocal lattice points, so that the intensity is known at these intermediate positions, thus increasing the sampling density. This is illustrated in Fig. 8, which shows the additional positions in reciprocal space where the continuous transform is known, for a hypothetical two-dimensional crystal containing a molecule that has a threefold noncrystallographic rotation axis.

In fiber diffraction there are some similarities to the optical case since the intensity is measured continuously along the layer lines, although this is only in one dimension. The use of continuity in assessing phases from isomorphous replacement is similar to the use of analyticity in optical phase retrieval. The increase in information content owing to the intensity's being recorded continuously in \(R\) in fiber diffraction is usually (but not always) more than offset by the loss of information owing to cylindrical averaging.\textsuperscript{71} Solvent leveling in fiber diffraction is more powerful than in crystallography since there is no lattice in the \(x\)-\(y\) plane, and so it is as strong as a support constraint in general imaging.

The above differences and similarities are of a fundamental theoretical nature. There are also a number of practical differences. The advantageous effects of atomicity in small-molecule crystallography and the possible presence of noncrystallographic symmetry in macromolecular crystallography are unlikely to occur in optical problems, although the equivalent of atomicity often occurs in astronomy. In crystallography one can often modify the specimen (isomorphous replacement) to obtain additional data, whereas this is usually not possible in optical problems such as astronomy, remote sensing, and nondestructive testing. In astronomical speckle imaging, however, one has the advantages that the images do contain some phase information and that a set of images with independent phase errors, with identical statistics, can be obtained. Another advantage in crystallography is that one knows what one is looking for; i.e., the image obtained has only to be good enough to fit the atomic model to. This means that larger phase errors can be tolerated than when one is reconstructing a general image.

**Comparison of Related Algorithms**

Despite the differences described above, there are a number of similarities between algorithms used in optical and crystallographic phase retrieval, with some of the optical techniques being motivated by methods used in crystallography. I will discuss first general-purpose phase-retrieval algorithms, followed by some of the astronomical techniques that are related to crystallographic methods.

If there is no phase information available, and no image constraints other than positivity, the only practical method for phase retrieval in general imaging is Fienup's algorithm. Other noniterative methods\textsuperscript{9,10} have been demonstrated in computer simulations, but, at present, they are too error sensitive to be used with real data. Fienup's error-reduction algorithm is seen to be identical to the density modification method in crystallography. Both techniques involve iterating between image space and Fourier space, enforcing known constraints in each. The actual constraints applied depend on course on the particular problem at hand, although positivity and extent (solvent leveling) are typically applied. Noncrystallographic symmetry is a useful constraint in this context. These methods are generally applicable and useful in both disciplines. The hybrid input-output algorithm has not been used in crystallography, although its use should be explored. However, the error-reduction algorithm usually reduces the phase error substantially in the first few cycles and so often may be sufficient in crystallography, in which larger phase errors can be tolerated.

The phase-extension procedure described in Section 6, in which phases at higher resolution are determined gradually in increasing-resolution shells, is necessary for convergence. It is interesting that precisely the same sort of procedure was found to be necessary by Fienup in order to reconstruct (difficult) complex images.\textsuperscript{89} In this case, the higher-resolution data were phased by gradually increasing the width of a weighting function in Fourier space as the iterations proceeded. It was necessary to start the iterations with a low-resolution image, which is equivalent to starting phase ex...
tension in crystallography with the low-resolution isomorphous phases. McCallum and Bates\(^9\) show that a similar technique can be used to accelerate convergence of Fienup's algorithm and is particularly effective for complex images. This also has some aspects in common with direct methods, in which the phases of a few reflections are propagated to a larger set, the phases of all reflections determined so far, being improved at each stage.

Baldwin and Warner\(^9\) describe the reconstruction of images of compact sources from amplitude-only radio astronomical interferometry measurements, which is motivated by crystallographic analogies. They begin by determining the sources by unraveling the autocorrelation, which is equivalent to solving a structure by inspection of the Patterson map. This is successful for either a simple image or one dominated by a few strong sources. Once the strong sources have been located, the remainder are located by using what are equivalent to difference Fourier maps in crystallography, to show sources that have been missed. The image is built up, using a sequence of such maps in which sources are added at each stage until all have been found.

As described above, although positivity is a common constraint in image space it is difficult to incorporate into Fourier space. The Karle–Hauptman determinants [relation (34)] do this, although in a somewhat complicated way. Following Karle and Hauptman, Cocke\(^6\) has used Cauchy's inequality to derive an optimized constraint on the Fourier magnitudes that ensures image positivity. This allows the positivity of the image to checked in Fourier space. A procedure that could be used to reduce negativity in the image, directly in Fourier space, would have obvious advantages.

Speckle holography is a simple idea that is clearly analogous to the heavy-atom method described in Section 5.

Both the Knox–Thompson and the triple-correlation phase-retrieval procedures described in Section 2 appear to be related to aspects of direct methods in crystallography. In each of these methods the phases are propagated from a small set of reflections to a much larger set. Although the correspondence is not exact, there are similarities between the phase-propagation equations in both Knox–Thompson and triple-correlation processing, and the triplet relationships (38) in direct methods. Referring to the description, following Eq. (14), of phase propagation in Knox–Thompson processing, denoting the phase of \(F(u)\) and \(F^*(u)F(u + \alpha)\) by \(\phi_u\) and \(\psi_{u,v-u}\), respectively, and letting \(v = u + \alpha\) allows the phase propagation to be written as

\[
\phi_v = \phi_u + \psi_{u,v-u},
\]

which bears a striking resemblance to relation (38). Turning now to bispectral processing, the phase-propagation Eq. (29) also resembles relation (38). This is not to say that the two situations are the same; the equations are different, and they arise from different causes: positivity and atomicity in one case and the fact that the cross correlation or bispectral phase is available as data in the other. It is interesting, however, that despite this the forms of the phase-propagation equations are quite similar. Phase propagation similar to this has been proposed in general imaging, but the phases thus produced are not particularly accurate.$^6\$\(^7\)$

Granrath\(^6\) has outlined an interesting proposal for using object positivity, in the form of Karle–Hauptman inequalities, to reduce the computational load in bispectral processing. Expanding the third-order Karle–Hauptman determinant [relation (34)] allows constraints to be placed on the phases of factors of the bispectrum Eq. (19) that depend on the parameter \(a\), given by

\[
\begin{aligned}
a &= \left[|F(u_1 + u_2)|^2 + |F(u_1)|^2 |F(u_2)|^2 - 1\right]/2|F(u_1)| |F(u_2)|, \\
&\times \left[2|F(u_1 + u_2)||F(u_1)||F(u_2)|\right],
\end{aligned}
\]

and that depend only on the magnitudes [that have been normalized by \(P(0)\)]. The strength of the constraint depends on \(a\), which therefore affects the variance of the bispectral phase. Granrath describes using the variances to assess the signal-to-noise ratio of the bispectral transfer function (although the bispectral transfer function is ideally positive, positivity is expected to be violated in practice with a finite number of specklegrams) at different locations in the bispectrum. This assessment may allow a small region of the bispectrum to be selected where the signal-to-noise ratio is high and the bispectrum to be calculated in this region only, for the purposes of phase determination, substantially reducing the amount of computation required.

Although in astronomical imaging one does not have the luxury of being able to modify the object as in crystallography, Rogers\(^6\) has pointed out that, if the field contains a variable star, then there is an automatic modification in the object. One can then apply processing similar to that in isomorphous replacement to restrict the possible phase values. Let the image be given by

\[
f(x, t) = a(t)\delta(x) + g(x),
\]

where \(a(t)\) denotes the time-dependent amplitude of the variable star and \(g(x)\) the remainder of the image. If measurements are made at two times, \(t_1\) and \(t_2\), in the cycle of the variable star (maximum and minimum brightness) and the autocorrelations calculated and subtracted, then

\[
A(x, t_1) - A(x, t_2) = [a^2(t_1) - a^2(t_2)]\delta(x) + [a(t_1) - a(t_2)]g(x) + g^*(-x),
\]

This allows the image plus its inverse to be reconstructed, a considerable simplification over the autocorrelation. This is a form of speckle holography in which the autocorrelation term \(g(x)\) is subtracted out, and the components \(g(x)\) and \(g^*(-x)\) are then distinct under less restrictive separation conditions.$^6$ This situation is analogous to SIR, in which the twofold ambiguity in the phase corresponds to the superposition of the image and its inverse.

Another astronomical technique that makes use of particular properties of the object is differential speckle interferometry, in which spectral variations across the object are used to obtain subdiffraction-limited resolution.$^6$\(^7\)$

**Potential Crossdisciplinary Applications**

Those working in a variety of imaging areas may see applications of some of the crystallographic algorithms described in this review to specific problems. I will therefore only outline here a few general areas in which applications of cross-disciplinary techniques may be useful.

The success of direct methods in crystallography indicates that it may be fruitful to explore adaptations to reconstructing general images (in astronomy, for example) that consist of sharp, isolated features. Modifications that incorporate...
the higher sampling density in imaging problems into these methods should further increase their power. It may also be useful to explore weighting schemes along the lines of those described in Section 6 to treat errors in the data in an optimum way. This could lead to an increase in the convergence rate of iterative algorithms.

On the crystallographic side, the use of Fienup's algorithms that incorporate negative feedback should be explored in density modification. This may assist convergence in difficult cases when there are minimal constraints (noncrystallographic symmetry, for example) in real space.

9. CONCLUDING REMARKS

Although crystallographic and optical phase problems have many characteristics in common, the presence of a crystalline lattice and the ability to modify the specimen in crystallography have led to the development of distinct techniques for phase retrieval. However, as has been shown here, despite these differences there are a number of commonalities in techniques. There is a continual need for investigators in one field to monitor advances in the other in case these prompt new applications.

Recent uniqueness results in optical phase retrieval are significant in that they lend confidence to results obtained from empirically inspired algorithms. More research is needed to understand uniqueness properties of crystallographic phase problems, however. The effects of atomicity and positivity on uniqueness or well-posedness in particular are not well understood. Iterative algorithms are widely applicable and can accommodate a variety of types of a priori information but are computationally intensive, and one always has to be concerned about convergence. If all the constraints could be mapped into one (image or Fourier) space, substantial computational savings could obviously be made. Direct methods in crystallography and algorithms based on zero sheets in optics do this, but more work is needed to make the latter less noise sensitive. More study is needed on methods of this type in general. It is to be expected that methods will be developed (as in speckle imaging) that take advantage of the special characteristics of particular problems leading to powerful, but restricted, algorithms. It is interesting that algorithms for general (less constrained) problems in both disciplines are iterative, whereas those that apply to more-restricted problems, again in both disciplines, involve similar bootstrapping procedures (that are noniterative) from a small initial phase set.

ACKNOWLEDGMENTS

My interest in phase retrieval has come about through interactions with many people. I am particularly grateful to Richard Bates of the University of Canterbury, New Zealand, through my association with whom I first became interested in the phase problem, as well as Richard Fright and Richard Lane with whom I had numerous discussions. I also benefitted from many discussions with members (past and present) of the molecular structure group at Purdue University and with other members of the crystallographic and fiber diffraction communities, including David Sayre. The reviewers offered many suggestions that substantially improved an earlier version of this paper. I am grateful to Jim Fienup for encouraging me to prepare this review for the feature issue of the Journal of the Optical Society of America. Finally, I thank the National Science Foundation for support (contract DMB-8606942), Deb Zerth for word processing, and Mark O'Neil for preparing some of the figures.

REFERENCES


