This work was supported by the National Science Foundation through Northwestern University Materials Research Center, Grant No. DMR 85-20280.

References

BETHE, H. (1928). *Ann. Phys. (Leipzig)*, **87**, 55-129.
BRITZE, K. & MEYER-EHMSEN, G. (1978). *Surf. Sci.* **77**, 131-141.
COLELLA, R. (1972). *Acta Cryst.* A**28**, 11-15.
COWLEY, J. M. & PENG, L. M. (1985). *Ultramicroscopy*, **16**, 59-68.
DAIMON, H. & INO, S. (1985). *Surf. Sci.* **164**, 320-326.
FESHBACH, H. (1958). *Ann. Phys. (Leipzig)*, **5**, 357-390.
FUJIMOTO, F. (1959). *J. Phys. Soc. Jpn*, **14**, 1558-1568.
ICHIMIYA, A. (1983). *Jpn. J. Appl. Phys.* **22**, 176-180.
INO, S. (1977). *J. Appl. Phys. Jpn*, **16**, 891-908.
INO, S. (1980). *J. Appl. Phys. Jpn*, **19**, 1277-1290.
KATO, N. (1952). *J. Phys. Soc. Jpn*, **7**, 397-414.
KIKUCHI, S. & NAKAGAWA, S. (1933). *Sci. Pap. Inst. Phys. Chem. Res. Tokyo*, **21**, 256-265.
KOHRA, K., MOLIERE, K., NAKANO, S. & ARIYAMA, M. (1962). *J. Phys. Soc. Jpn*, **17**, Suppl. B-II, 82-85.
MA, Y. (1990). *Acta Cryst.* submitted.
MA, Y. & MARKS, L. D. (1989). *Acta Cryst.* A**45**, 174-182.
MA, Y. & MARKS, L. D. (1990). *Acta Cryst.* A**46**, 11-32; erratum: *Acta Cryst.* (1990). A**46**, 626.
MCRAE, E. (1966). *J. Chem. Phys.* **45**, 3258-3276.
MCRAE, E. A. & CALLDWELL, C. W. (1976). *Surf. Sci.* **57**, 63-76.
MCRAE, E. G. (1979). *Rev. Mod. Phys.* **5**, 357-390.
MAKSYM, P. A. & BEEBY, J. L. (1981). *Surf. Sci.* **110**, 423-438.
MAKSYM, P. A. & BEEBY, J. L. (1982). *Appl. Surf. Sci.* **11/12**, 663-676.
MARKS, L. D. & MA, Y. (1988). *Acta Cryst.* A**44**, 392-393.
MARKS, L. D. & MA, Y. (1989*a*). *Ultramicroscopy*, **29**, 183-191.
MARKS, L. D. & MA, Y. (1989*b*). *Ultramicroscopy*, **31**, 241-244.
MENADUE, J. F. (1972). *Acta Cryst.* A**28**, 1-11.
MIYAKE, S. & HAYAKAWA, K. (1966). *J. Phys. Soc. Jpn*, **21**, 363-378.
MIYAKE, S., KOHRA, K. & TAKAGI, M. (1954). *Acta Cryst.* **7**, 393-401.
MOON, A. R. (1972). *Z. Naturforsh. Teil A*, **27**, 390-395.
PENG, P. M. & COWLEY, J. M. (1986). *Acta Cryst.* A**42**, 545-552.
PENG, P. M. & COWLEY, J. M. (1988). *Surf. Sci.* **199**, 609-622.
SIEGEL, B. M. & MENADUE, J. F. (1967). *Surf. Sci.* **8**, 206-216.
SOMORJAI, G. A. (1981). *Chemistry in Two Dimensions, Surfaces*, 1st ed., p. 126. Ithaca: Cornell Univ. Press.
TURNER, P. S. & COWLEY, J. M. (1981). *Ultramicroscopy*, **6**, 125.
WANG, Z. L., LU, P. & COWLEY, J. M. (1987). *Ultramicroscopy*, **23**, 205-222.
WHELAN, M. J. & HIRSH, P. B. (1957). *Philos. Mag.* **2**, 1121-1142, 1303-1324.
ZHAO, T. C., POON, H. C. & TONG, S. Y. (1988). *Phys. Rev. B*, **38**, 1172-1182.
ZHAO, T. C. & TONG, S. Y. (1988). *Ultramicroscopy*, **26**, 151-160.

Acta Cryst. (1990). A**46**, 606-619

Direct Solution of Continuous Densities Given the Fourier Magnitudes

BY ROBERT W. HARRISON

Crystallography Laboratory, National Cancer Institute - FCRF, PO Box B, Frederick, MD 21701, *USA*

(*Received* 12 *September* 1989; *accepted* 12 *February* 1990)

Abstract

In order to apply direct methods routinely to macromolecular crystals it will be necessary to generate a non-atomic theory which is applicable to continuous densities. Reformulation of the 'phase problem' in terms of deconvoluting an autocorrelation function or Patterson synthesis reduces the problem from a theoretically intractable transcendental problem to a system of simultaneous quadratic equations. These quadratic equations may, in principle, always be solved by conjugate direction search techniques. The phase problem is shown to be a class P problem, admitting a deterministic solution in polynomial time. Two algorithms are presented with running times proportional to N^2_{points} and $N_{\text{points}} \log N_{\text{points}}$ per step. These algorithms are a pixel-by-pixel search and a conjugate gradients search. When the data are exact and complete the Fourier magnitudes are readily inverted by them to find the image. An example with real data, from a 15mer of DNA, is also shown.

Introduction

The phase problem in crystallography occurs because it is only possible, in general, to measure the amplitude of the diffracted X-ray. A great deal of effort has gone into various methods for overcoming this problem. Experimental means like isomorphous replacement and anomalous scattering (Blow & Crick, 1959; Watenpaugh, 1985; Hendrickson, Smith & Sheriff 1985) for macromolecules, as well as theoretical approaches like direct methods (Karle & Hauptman, 1950; Ladd & Palmer, 1980) for small molecules have been developed to find useable values for the phases. These have been successful, but the general solution to the phase problem has remained elusive. In particular, it has been very difficult to invert numerically the observed magnitudes and obtain the electron density when the density neither is composed of a small number of atoms, nor diffracts to high resolution. This is typically the case with macromolecular structures.

0108-7673/90/070606-14$03.00

The central computational question in algorithm design is the complexity of the algorithm. For example, one could try all possible combinations of phases, and some of these would be correct enough to resolve the structure. The problem is that this literally could take forever. Generally, algorithms can be divided into two classes, P and NP. If the running time is proportionate to some finite polynomial of the number of unknowns the algorithm is a member of class P and runs in polynomial time. Otherwise, it runs in non-polynomial time and is a member of class NP. Class P is a desirable property because the run time of the program increases slowly with the number of unknowns. The only improvement over a class P algorithm with provable convergence is an analytical expression. It would be of interest to find an algorithm in class P which determines phases.

Classically, the phase problem has been described as that of finding a positive electron density whose Fourier transform has the observed intensities. If the structure factors are normalized the problem may be stated in terms of atomic positions. Either way, this problem requires the solution of a large number of simultaneous transcendental equations. Direct methods attempt to do this by analytic expressions related to the probability of the phase of a reflection or of an invariant sum of phases (Hauptman, 1972; Karle, 1982). Direct iterative methods like the maximum-entropy method have had limited success at the solution of the transcendental system of equations (Harrison, 1989; Bricogne, 1984; Marvin, Bryan & Nave, 1987). Experimentally, measures of the change in the intensities of the structure factors when the structure is perturbed in a limited manner with a heavy atom have been used to deduce phases.

The conventional direct approaches often work, and for small molecules do work very well. Computationally, they try to solve a combinatorial optimization problem, which does not have an efficient deterministic solution (Ladd & Palmer, 1980). The complexity is at least exponential, placing them squarely in class NP. This arises because the optimization problem is one of simultaneous transcendental equations. Not only do these transcendental equations lack a rational inverse, they have multiple minima and multiple stationary points at non-minima.

While the phase and magnitude representation of the scattered X-ray is physically and intuitively appealing it is not the only way that the scattering problem and its inverse can be treated. The intensities are a direct measure of the autocorrelation function or Patterson function of the electron density. The essential problem of crystallography is to find a bounded and physically realistic electron density which reproduces the observed autocorrelation function. It is not required that this density be found *via* the Fourier magnitude and phase representation. It is shown that the autocorrelation function may be written as a multidimensional quadratic problem which is, in principle, much easier to solve than the traditional transcendental one.

In this paper three major points are developed. First, the failure of linear models of image degradation to produce new phase information is demonstrated. This is not surprising given the orthogonality of the terms in the Fourier transform. Second, the difficulties in the construction of a conventional technique are sketched. In particular, any method which only uses, either explicitly or implicitly, first derivatives or secants with respect to the phase as a measure of phase error will fail to resolve a mixture of enantiomorphs. Third, techniques for deconvoluting a Patterson are developed. This approach is a dual, or reformulated version, of the traditional phase problem. It is a dual, because to solve the Patterson and find the image (and from that the phases), it is not necessary to use the phase variables or leave the set of real numbers. Deterministic algorithms with absolute convergence in polynomial time can be constructed for deconvolution of the Patterson function. These algorithms can be implemented on both current serial computers and future parallel architectures. When the data are complete and exact, as in a two-dimensional test model, these algorithms recover the phases from the Fourier magnitudes. Some of the numerical issues involved in the practical implementation of these algorithms are developed. Although these have not all been resolved yet, some results are shown.

The algebra of circulant matrices

The mathematical development of this approach will depend on the properties of a class of matrices known as circulant matrices (Davis, 1979; Gonzalez & Wintz, 1977). Because the eigenvectors of circulant matrices are the discrete Fourier transform, the algebra of these matrices forms an effective tool for the manipulation of Fourier transforms. Since the representation of an electron density map on a computer is always a discrete Fourier transform rather than a Fourier integral it is useful to develop the theory in terms of the computer representation. This section describes the basic properties of these matrices and their close relatives, circulant-block-block-circulant matrices.

Definition

A circulant matrix is a square matrix where all of the elements are present on any one row or column and each row is related to the next by a unit shift in index. A circulant-block-block-circulant (CBBC) is a matrix made of block circulant matrices where the block matrices are also arranged in circulant fashion. Such a matrix arises in two-dimensional problems.

Each row in a CBBC matrix is related to the two dimensions in the same way as a two-dimensional array is indexed in a computer language like Fortran. CBBC matrices may further be arranged in circulant fashion to form the equivalent matrix for a three-dimensional problem. Since these matrices have the same addition, multiplication and inverse properties they will all be referred to as circulants and the dimensionality will depend on the context. Since all of the rows in these matrices are related to the other rows by a permutation, only the storage needed to keep one row in memory is needed for a computer program.

Eigenvectors

Circulant matrices and their two- and three-dimensional generalizations have one major property which makes them useful. The eigenvectors of any circulant are equivalent to the Fourier transform (Davis, 1979; Gonzalez & Wintz, 1977). The eigenvalues of the circulant can be found by the Fourier transform of the first row of the circulant:

$$C_{nn} = F^*_{nn} \Lambda_{nn} F_{nn}$$

where C_{nn} is an n by n circulant, F_{nn} is the Fourier transform, and Λ_{nn} is the n by n diagonal matrix of the eigenvalues.

F_{nn} is a matrix which when multiplied by a vector results in the Fourier transform of the vector. This matrix is a symmetric unitary matrix and its inverse is simply its complex conjugate. The fast Fourier transform works by taking advantage of the structure of the matrix to find ways of taking the matrix product with fewer than n^2 multiplications (Winograd, 1978). F_{22} and F_{44} are shown below.

$$F_{22} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$F_{44} = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}.$$

Addition, multiplication and pseudo-inverses

Addition of circulants is done in the same way as other matrices, each pair of elements being added to produce the corresponding element in the output. Since all the unique elements of a circulant are in one row or column, this need only be done for one row or column of the matrices. Multiplication could be done in the same way as usual, but it is possible to speed it up greatly by the use of the fast Fourier transform (FFT) (Davis, 1979; Gonzalez & Wintz, 1977; Winograd, 1978). The circulants are first reduced to their eigenvalues, these are multiplied, and then the product is reconstructed with the inverse Fourier transform.

$$C^1_{nn} C^2_{nn} = F^*_{nn} \Lambda^1_{nn} F_{nn} F^*_{nn} \Lambda^2_{nn} F_{nn}$$
$$= F^*_{nn} \Lambda^1_{nn} \Lambda^2_{nn} F_{nn}.$$

A matrix and an arbitrary vector may be quickly multiplied together by forming the circulant matrix with one row equal to the vector and using the FFT to multiply the two matrices. It should be noted that the above equations are also known as the convolution theorem and multiplication by a circulant results in a convolution.

Since it is easy to find the eigenvalues of a circulant matrix, the inverse and Moore-Penrose pseudo-inverse may be found (Davis, 1979). From the equation above, if Λ^1 has elements which are the inverse of the elements of Λ^2 then C^1 and C^2 are inverses. The inverse of a circulant is found by first calculating the eigenvalues, and then using the inverses of these to generate a new circulant. The pseudo-inverse is constructed by leaving as zero any of the elements of Λ which are zero, but using the inverse of the elements when they are not. The pseudo-inverse of a matrix C will be referred to as C^+. When all the eigenvalues are non-zero then pseudo-inverse and inverse are identical, but the pseudo-inverse will be used as it is defined when some eigenvalues are zero (preventing calculation of an inverse), and automatically gives the least-squares solution in those cases.

Linear image degradation models

The conventional image processing problem will be described to show how circulants may be used. In image processing, an observed copy of a degraded image is restored by making a model of the degradation process, and then using this model to enhance the image. The linear model for degradation will be developed and extended to the problem of missing phase information. The inability of the linear model to restore missing phase information will be shown.

The degradation of an image can usually be described with a linear model. The observed image ρ_{obs} and the true image ρ_{ideal} are related *via* the linear equation

$$\rho_{\text{obs}} = A\rho_{\text{ideal}} + n$$

where A is a matrix (the degradation matrix) which describes the systematic distortion of the image and n is a vector which represents the added noise. Provided that the systematic distortion is nearly spatially stationary (*i.e.* the same throughout the image), then A is to a good approximation a circulant matrix. For crystallographic problems A will always be circulant

because of the lattice periodicity. A crystallographic example of this is the thermal factor; the electron density of the atoms is broadened by the convolution with a Gaussian factor which represents the average motion of those atoms.

Since A is circulant, $A^{\dagger}$ may easily be calculated. This can then be used to find an estimate for ρ_{ideal},

$$\rho_{\text{ideal}} \simeq A^{\dagger}\rho_{\text{obs}}$$

with an error bound of the magnitude $\|A^{\dagger}n\|$ for a norm $\|\ \|$ (a norm is a measure of the size of a matrix or vector; it may be thought of as an absolute value). This algorithm is known as inverse filtering (Gonzalez & Wintz, 1977). It is notorious for its sensitivity to noise. The norm of $A^{\dagger}$ can be large so that even small amounts of noise result in large errors in the recovered image. In practice the simple pseudo-inverse described here is modified so that its norm is reduced. Instead of inverting all non-zero eigenvalues, those which are 'small' by some definition are left untouched or set to one. Depending on the exact procedure this produces one of a whole family of quasi-inverses such as the Wiener filter (Gonzalez & Wintz, 1977).

The linear model is insufficient for missing reflections

Extending the linear model to include degradation due to missing reflections may be done in a simple manner. However, even in the best case (the absence of noise), this model is not useful since the degradation matrix produced is its own pseudo-inverse.

Construction of the degradation matrix starts from its eigenvalues. These are either 1 when the phase and magnitude are known or 0 when they are not. The degradation matrix is then

$$D = F^{*}\Lambda_{d}F.$$

Since this construction parallels Dirichlet's integral in Fourier integral theory, it will be referred to as the Dirichlet matrix (Phillips, 1984). The surprising properties of this matrix are summarized below:

$$\begin{aligned} D &= (D)^{n}, \text{ where } n \text{ is any positive integer} \\ D^{\dagger} &= D, \ D \text{ is its own pseudo-inverse} \\ D^{\dagger}D &= DD^{\dagger} = D. \end{aligned}$$

Any circulant matrix, A, is a pseudo-inverse to D provided that whenever D has an eigenvalue of 1, A does as well.

These properties mean that any phase determination algorithm based on a linear model for the degradation due to missing data will not give reliable estimates for the unknown phases. Using the linear solution, and assuming $A^{\dagger}$ is any linear pseudo-inverse (which could be $D^{\dagger}$), we obtain

$$\begin{aligned} \rho_{\text{obs}} &= D\rho_{\text{ideal}} \\ A^{\dagger}\rho_{\text{obs}} &= A^{\dagger}D\rho_{\text{ideal}} \\ A^{\dagger}\rho_{\text{obs}} &= F^{*}\Lambda_{A}^{\dagger}\Lambda_{D}F\rho_{\text{ideal}} \\ A^{\dagger}\rho_{\text{obs}} &= F^{*}\Lambda_{D}F\rho_{\text{ideal}} \\ A^{\dagger}\rho_{\text{obs}} &= D\rho_{\text{ideal}} \\ A^{\dagger}\rho_{\text{obs}} &= \rho_{\text{obs}}. \end{aligned}$$

Since the eigenvalue of 0 in D, corresponding to a missing reflection, will set the eigenvalue of the product with $A^{\dagger}$ to 0 no new phase information can be found by the pseudo-inverse. Because a new phase cannot be created by pseudo-inverses, a value must be chosen independently from the restoration problem. This chosen phase then has to be refined in some way to produce an accurate value.

Eigenvalue iteration algorithms

The degradation matrix due to phase errors, as opposed to missing phases, can be written in a similar fashion to the Dirichlet matrix. The important difference is that the matrix does not have constant coefficients; instead the eigenvalues depend on the current phases. While the linear formalism can be preserved, in that a degradation matrix is used, the problem is non-linear because the degradation matrix is not constant.

The eigenvalues of the degradation matrix for phase error are $e^{i\varphi_{\text{current}}}/e^{i\varphi_{\text{true}}}$. However, as the true values for the phases are unknown, it is not possible to construct an exact pseudo-inverse. Instead, an approximate pseudo-inverse is calculated by comparing the image with some physical criterion and estimating the phase difference. This phase difference is used to estimate the new phase, and the process is iterated until convergence. Crystallographic criteria include positivity, atomicity, connectedness, flat solvent regions, non-crystallographic symmetry, depending on the type of problem (Wang, 1985; Tulinsky, 1985). Statistical ideas such as maximum entropy can be used to perform this comparison.

Most of the direct methods and density modification methods used in crystallography are variants on this theme.

Limits to eigenvalue iteration

Clearly, the convergence of an eigenvalue iteration algorithm will depend on the accuracy with which the phase error degradation matrix has been estimated. In order to show convergence it is necessary to show that a convergent sequence of estimates can be made.

Forming a chain of increasingly reliable phase estimates can be difficult. The problem is that there are multiple stationary points (places where the first derivative with respect to the phase is zero) in any application of a map improvement operator. Because of the mean value theorem, if this operator is not a constant, it must possess at least a maximum and a minimum over the unique range of the phase variable. The derivative will be zero at those points so that the phase shift on application of the operator will be zero there as well. Because the electron density ρ is real, it is possible to find a precise relationship between these stationary points.

To illustrate, let the continuous operator $O(\rho)$ be an operator which generates a better map, for example $O(\rho)=\rho^2$ (Sayre's equation) for small molecules. If the sequence defined by iterative application of $O(\rho)$ converges, at the convergence point the phases of $O(\rho)$ and ρ must be the same. $O(\rho)$ can always be defined as an operator which restores the observed magnitudes. This modified operator will only have derivatives in the phase variable.

$$O'(\rho)=F_{nn}^{*}\Lambda_r F_{nn}O(\mathcal{P})$$

where $\mathcal{P}$ is the density, ρ, in circulant form, F_{nn} is the Fourier transform operator, and

$$\Lambda_r=|F_{\rm obs}|/|F_{nn}O(\mathcal{P})|$$

is a diagonal matrix restoring the observed data $F_{\rm obs}$.

In these circumstances $\mathrm{d}[O'(\rho)-\rho]/\mathrm{d}\varphi$ and $\mathrm{d}[O'(\rho)-\rho]/\mathrm{d}|F_{\rm obs}|$ equal zero at the convergence point. Since the Friedel mate, $\bar{\varphi}$, is simply $-\varphi$, the derivative $\mathrm{d}[O'(\rho)-\rho]/\mathrm{d}\bar{\varphi}$ is also zero. The enantiomorphic phase in $P1$ is simply generated by swapping phases with their Friedel mates. Therefore the derivative with respect to the enantiomorph phase, $\mathrm{d}[O'(\rho)-\rho]/\mathrm{d}\varphi_{\rm enan}$, is zero when the derivative with respect to the phase is zero. This shows that methods which only use first-derivative information, or secant information, will fail to resolve a mixture of enantiomorphs. The secant is used implicitly if the phase shift after the application of a density modification is used as a measure of correctness.

This does not demonstrate that all phase search algorithms converge incorrectly; rather, it shows that the simplest fixed-point iteration, the naive application of a function to ρ and extraction of the phase shift as a measure of error, can converge incorrectly. This may be remedied by using higher-order derivative information, or a direct search over the phase variable to ensure that the operation really has converged to its best value.

Pseudo-inverse algorithm

The autocorrelation function or Patterson may be thought of as the image convoluted with itself (or its enantiomorph depending on how convolution is defined). In the linear model described above the image is its own degradation matrix. If $\mathcal{P}$ is a circulant of the image ρ and P is the Patterson then

$$P=\mathcal{P}\rho.$$

Finding $\mathcal{P}^{+}$ for some approximate density makes it possible to find the solution for ρ such that ρ and the approximate density convolute to give the observed Patterson

$$\rho=\mathcal{P}^{+}P$$

Transference of this to the eigenvalues gives rise to the β synthesis (Ramachandran & Raman, 1959; Main, 1979)

$$F_\beta=(|F_{\rm obs}||F_{\rm obs}|/F_{\rm calc})^{*}$$

where F is now the Fourier coefficient.

Comparison of ρ with a physical criterion and recycling through this procedure defines a fixed-point iteration which might converge to an invariant value for ρ. A fixed-point iteration has a small convergence radius, but the accuracy and speed if it converges will be high. A linear analogy of this process is the iterative improvement of the solution to a linear equation (Conte & de Boor, 1980).

This construction may also be used to generate the Newton minimizer for the square error in the Patterson. As shown below the square error in the Patterson, $\int \mathrm{d}V\,(P_{\rm obs}-P_{\rm calc})^2$ where $P_{\rm obs}$ is the Patterson calculated with the observed magnitude data $F_{\rm obs}$ and $P_{\rm calc}$ is calculated from the calculated magnitudes $|F_{\rm calc}|$, is equivalent to the sum $\sum(|F_{\rm obs}|^2-F_{\rm calc}^{*}F_{\rm calc})^2$. The first derivative of this sum is simply $-2(|F_{\rm obs}|^2-F_{\rm calc}^{*}F_{\rm calc})F_{\rm calc}^{*}$. Dividing the residual by the negative of the derivative gives the Newton minimizer,

$$\sum(|F_{\rm obs}|^2-F_{\rm calc}^{*}F_{\rm calc})/2F_{\rm calc}^{*},$$

which is just half the difference between the current map and its β synthesis.

An all-real algorithm with global convergence

Writing the Patterson function as a real-space convolution, $P=\mathcal{P}\rho$, shows that it may be regarded as a quadratic function. This is especially clear if one pixel at a time is examined. The calculated Patterson may be written as

$$\begin{aligned}P_{000}&=\sum\rho^2\\ \vdots&=\vdots\\ P_{i-i'j-j'k-k'}&=\sum\rho_{ijk}\rho_{i'j'k'}\\ \vdots&=\vdots\end{aligned}$$

which is quadratic. When $\rho>0$ the system of equations is also a positive definite system of quadratic equations. If the squares of the difference between the calculated and observed Pattersons are

summed the resulting equation is quartic. This quartic function has either a unique minimum, a single solution, or a pair of homometric solutions as a function of any one pixel when the density value of the pixel is constrained to be bounded (positive) and real. The function is also quartic in the step size for a constant search direction (*i.e.* for α in $\rho_{new} = \rho_{old} + \alpha\rho_{search}$). Minimizing the quartic function, $\sum (P_{calc} - P_{obs})^2$, over a bounded (*i.e.* positive) density will result in a solution for the density which is at worst homometric to the true density. This system of equations is completely determined because there are equal numbers of variables and observations.

These equations could be solved by trial and error with simulated annealing (Corana, Marchesi, Martini & Ridella, 1987). However, it is possible to derive a numerical equation solver with global convergence for this problem. The general algorithm can now be described. From some starting point, either a partial structure or a random starting set, a minimization direction is chosen. This direction may be either a pixel or a gradient direction. The step along this search direction which minimizes the square error in the Patterson function while meeting image constraints like positivity is found. The density is now updated and a new minimization direction is chosen. This direction is chosen to be conjugate (orthogonal when the second derivative is used as a metric) to all of the previously chosen directions and the process repeats from the search for the minimum along this direction. Since the Patterson may be calculated by the convolution of the image with itself, this algorithm could find the image and the phases without ever leaving the set of real numbers. The details of two practical versions of this algorithm are presented below.

The ability to find the global optimum along each conjugate search direction is a sufficient condition for a conjugate direction search over a bounded domain to converge globally (McCormick, 1983). Since the conjugate direction algorithm converges in polynomial time when this condition is met, the phase problem is a member of class P and there exists a determinative method to solve it. If a conjugate direction line search can find the unique minimum or first solution of the square error between the calculated and observed Pattersons then it will be globally convergent.

Homometric solutions

This algorithm is guaranteed to find a solution. The only problem is that the solution found may not be the desired one. It is possible to have homometric solutions, or multiple densities which have the same Patterson.

The way these arise is especially easy to see in one dimension. Any one-dimensional sampled function may be fit by an interpolating polynomial. So may its Fourier transform (Isrealevitz & Lim, 1987; Lane, Fright & Bates, 1987). (Note that an Nth-order polynomial is uniquely determined by its N zeros, or equivalently by its value at $N+1$ samples.) The fundamental theorem of algebra guarantees the ability to factor any one-dimensional polynomial (Isrealevitz & Lim, 1987), so this polynomial may always be factored into N complex factors. The interpolating polynomial for the intensities of the transform may be found by taking the product of the interpolating polynomial and its complex conjugate.

$$F(\rho) = e^{i\alpha}(x - x_0)(x - x_1)\ldots(x - x_{n-1})$$

$$F^*(\rho) = e^{-i\alpha}(x - x_0^*)(x - x_1^*)\ldots(x - x_{n-1}^*)$$

$$I = (x - x_0^*)(x - x_0)\ldots(x - x_{n-1}^*)(x - x_{n-1})$$

where α is an arbitrary constant due to origin choice, x is a real number corresponding to distance in reciprocal space, and x_i are the, possibly complex, zeros. Any one-dimensional density whose Fourier transform is formed by taking one factor of each complex pair will have the same Patterson as the original density. In the absence of a physical criterion like positivity, these 2^n solutions are indistinguishable. Any centrosymmetric function, with all real zeros in the Fourier transform, is uniquely determined because the factors of $F(\rho)$ and $F^*(\rho)$ are the same. It should be pointed out that with a positivity criterion even fairly complicated non-centrosymmetric densities can be solved; however, the solution may not be unique.

This analysis may make it seem impossible to solve anything in higher dimensions as the number and complexity of the zeros will increase unboundedly. However, because there is no fundamental theorem of algebra in more than one dimension, the two- and three-dimensional problems are easier than the one-dimensional one. A two-dimensional polynomial may or may not be factorable and in general is not (Hayes, 1987; Isrealevitz & Lim, 1987; Lane, Fright & Bates, 1987). Therefore there will be far fewer indistinguishable, solutions than in the one-dimensional case. As in the one-dimensional case, solving for the phases is equivalent to factoring the interpolating polynomial of the intensities. The polynomial of the intensities is always factorable at least once because the intensities are the product of F and F^*. If the interpolating polynomial of the Fourier transform is further factorable then several densities exist which will have the same Patterson (this is not counting trivial differences like origin and enantiomorph). In this case positivity and chemical information are required to differentiate between the solutions.

Numerically, the situation is somewhat more complicated as the effects of sampling must be considered. This is especially important for crystallographic applications where the sampling is dictated by the

crystal lattice and is not the choice of the crystallographer. Basically, it is possible to sample two different densities so that the sampled densities produce the same Patterson. As an illustration, the two 4×4 images

1 1 0 0
1 1 0 0
1 1 0 0
0 0 1 1

and

1 0 0 1
1 1 0 0
1 0 0 1
0 0 1 1

have the same Patterson and the image retrieved by a solver is a matter of chance. However the 8×8 images

1 1 1 1 0 0 0 0
1 1 1 1 0 0 0 0
1 1 1 1 0 0 0 0
1 1 1 1 0 0 0 0
1 1 1 1 0 0 0 0
1 1 1 1 0 0 0 0
0 0 0 0 1 1 1 1
0 0 0 0 1 1 1 1

and

1 1 0 0 0 0 1 1
1 1 0 0 0 0 1 1
1 1 1 1 0 0 0 0
1 1 1 1 0 0 0 0
1 1 0 0 0 0 1 1
1 1 0 0 0 0 1 1
0 0 0 0 1 1 1 1
0 0 0 0 1 1 1 1

do not have equivalent Pattersons. However, they are *nearly* equivalent and will be referred to as near-homometric images. It should be pointed out that the two 4×4 images are both samplings of one 8×8 image, which illustrates the dependence of homometry on sampling.

The limit to the convergence of Patterson deconvolution is the existence of homometric solutions. These can either be a true characteristic of the density, or can arise due to the finite sampling of the Fourier transform. Most of the time a bound on the density, like positivity, will resolve these ambiguities. Images which are homometric or near-homometric can be difficult to solve. Changing the sampling or changing the Fourier coefficients included (either by resolution range or a thermal factor) may resolve this problem. In the worst case, where there are two positive images which agree with chemical knowledge, it may be necessary to make a phase-sensitive measurement like isomorphous replacement or anomalous scattering in order to bias the solution into one or the other image.

Reciprocal-space measure of error in the Patterson

So far the assumption has been made that the data have been completely observed. Unfortunately, this is not often the case with real problems. The observed Patterson is not just ρ convoluted with itself; ρ has also been degraded by the Dirichlet matrix. The effects of this convolution cannot be ignored. For example there may be no purely positive solution for ρ if unobserved reflections are treated as though they had zero intensity. The function $\int \mathrm{d}V\,(P_{\mathrm{calc}}-P_{\mathrm{obs}})^2$ becomes $\int \mathrm{d}V\,(P_{\mathrm{calc}}-DP_{\mathrm{obs}})^2$ when there is finite resolution. This function is expensive to calculate in real space because of the convolution with D.

It is possible to show that there is a sum in reciprocal space which is equivalent (within a constant multiplier and an additive constant) to the square error in the Patterson. The sum, $\sum (F^*_{\mathrm{calc}}F_{\mathrm{calc}}-|F_{\mathrm{obs}}|^2)^2$, has proportionate derivatives and the same zeros as the square error in the Patterson.

$$\frac{\mathrm{d}}{\mathrm{d}F^*_{\mathrm{calc}}F_{\mathrm{calc}}}\int \mathrm{d}V\,(P_{\mathrm{calc}}-P_{\mathrm{obs}})^2$$

$$=\int \mathrm{d}V\frac{\mathrm{d}}{\mathrm{d}F^*_{\mathrm{calc}}F_{\mathrm{calc}}}(P_{\mathrm{calc}}-P_{\mathrm{obs}})^2$$

$$=\int \mathrm{d}V\,2(P_{\mathrm{calc}}-P_{\mathrm{obs}})\cos(2\pi\mathbf{h}\,.\,\mathbf{x})$$

$$=4\pi(F^*_{\mathrm{calc}}F_{\mathrm{calc}}-|F_{\mathrm{obs}}|^2)$$

$$\frac{\mathrm{d}}{\mathrm{d}F^*_{\mathrm{calc}}F_{\mathrm{calc}}}\sum (F^*_{\mathrm{calc}}F_{\mathrm{calc}}-|F_{\mathrm{obs}}|^2)^2$$

$$=2(F^*_{\mathrm{calc}}F_{\mathrm{calc}}-|F_{\mathrm{obs}}|^2).$$

The chain rule may then be used to show the equivalence of any particular derivative of F_{calc}. Because this sum is in the eigenspace of D, the evaluation of the function for finite resolution is trivial.

This result allows the realistic definition of the degree to which the equations are determined. Clearly, if there are as many reflections as pixels the problem is completely determined. Practically, the problem will always be underdetermined because of both experimental errors, and the need to oversample in order to avoid aliasing (Ten Eyck, 1977).

Practical algorithms

Having described a theoretical algorithm for the phase problem, it is important to address some of the issues involved in the construction of a computer program which will implement the algorithm. Two implementations will be described. The first of these is a pixel-based search. It is relatively expensive, but could be implemented at high efficiency on a parallel processor. The second is a conjugate gradients search. This approach is faster than the pixel-based approach,

and the implementation can be written in a vectorizable fashion.

Efficient pixel search algorithms

The simplest way to approximate a conjugate shift algorithm is to use orthogonal shifts. This may be done by the use of a pixel-by-pixel search. Such an algorithm strongly resembles the Gauss–Siedel relaxation scheme often used for linear systems of equations in that the error as a function of one degree of freedom at a time is minimized, and each new minimum is immediately used for the next search. On conventional serial or vector processors the cost of this algorithm is considerably higher than that of the general direction algorithm described below, but on a sufficiently large parallel machine this sort of approach could be implemented with high efficiency. For this reason it is useful to describe some of the numerical issues involved in the efficient implementation of a pixel-based search.

A useful theorem of Lagrange

The key to efficient algorithms is the speed with which the error as a function of step size along a search direction can be found. Lagrange showed that the polynomial which interpolates a set of points is unique (Conte & de Boor, 1980). That is, if samples from a finite-order polynomial are interpolated by another finite-order polynomial of the same degree, then the original and the interpolating polynomial must be the same function. The error in the autocorrelation function is a quadratic polynomial for constant search direction. The square error is quartic. The practical result from this is that, for fixed cost, the error along the search direction can be exactly parameterized by a quartic polynomial to the limits of the numerical accuracy of the computer. This accuracy may be optimized by the use of Chebyshev zeros (Conte & de Boor, 1980). The cost to minimize a low-order polynomial is very small, especially compared with the cost of evaluating the error in the Patterson.

The pixel-based algorithm

Efficient pixel algorithms utilize both quartic interpolation and precomputation. Since the search direction is small in comparison to the map as a whole, care must also be taken to minimize numerical error in evaluating the difference in the error as a function of the step length. Following a general Gauss–Siedel recipe, each pixel in the asymmetric unit is adjusted to minimize the error. The Fourier transform of a pixel is simply $e^{2\pi ihx}$ which may be expanded in terms of $e^{2\pi ix}$ and evaluated in $O(N)$ complex multiplications. Given the current Fourier transform of the image the new Fourier transform will be $F_{\text{current}} + (\alpha - \rho_{\text{pixel}})F_{\text{pixel}}$. The error in the Patterson is then evaluated in reciprocal space. It is important to sum the differences in the error rather than the error because the Fourier transform of each pixel is small compared with the whole map and floating-point errors will inhibit convergence. It is easy to evaluate the pixel transform under a crystallographic symmetry, but it is important only to search pixels within an asymmetric unit if this is done. This process is repeated several times, until the error is low, and then a combination of the minimum cross entropy approach (Harrison, 1989) and the beta synthesis is used to find a solution. If the process fails, pixel-by-pixel minimization is restarted and run to a lower error. The cost per pixel is $O(N)$ and there are N pixels giving a total complexity of $O(N^2)$ per optimization step over the whole map.

Values for the difference in the error are found for five steps, and these are fit with the interpolating polynomial. The steps should be chosen as the Chebyshev points bracketing a zero step $\{\cos[\pi I/(N_{\text{step}}-1)]$ with $I = 0, \ldots, N_{\text{step}}-1\}$. Since one of the Chebyshev points is zero for five points, the difference in error for one step is known without any work. Newton divided differences are used to find the polynomial (Conte & de Boor, 1980), which is then minimized by a quadratic search. Since a small number of extremely large peaks is nearly homometric to almost anything, it is advantageous to limit positive shifts to less than F_{000}/V_{cell}. Negative shifts need only be bound by positivity.

A general search direction algorithm

The algorithm presented in the previous section is primarily of theoretical interest as large parallel computers are rare, and the architectures and operating systems of those machines which are available are in a state of flux. Fast scalar and vector machines are common, and this section describes an algorithm which may be efficiently implemented on one of these. The conjugate gradients algorithm has already been described in detail in the literature (Powell, 1977; McCormick, 1983), so it is appropriate to concentrate on the specific details of a practical implementation of this algorithm to Patterson solution. The search direction is allowed to be an arbitrary vector, in this case the steepest descent direction, and the search directions are forced to be conjugate with a recursive formula. This algorithm is given in flow-chart form in Fig. 1.

The central problem in the design of a general direction algorithm is how to enforce bounds on the electron density. With the single-pixel approach this is trivial: when the pixel value hits the bound, stop. In the case of a general direction search, the optimum step length may put some pixels outside the bounds. These could be simply truncated, or the step size

reduced to the largest step with all of the pixels in bounds, but neither of these alternatives is adequate.

The optimal solution to this problem is to transform the search space into another space where the constraints are met automatically. For example, we might work in a space defined solely by the phases of the Fourier series; all trial solutions would have the same magnitudes. In maximum entropy methods the bound of positivity is usually maintained by working in a Lagrange multiplier space. By working in the λ space of e^{λ}, shifts are mapped from $-\infty$ to ∞ to the range 0 to ∞. In test cases it has been found useful to include an upper bound as well as a lower one. The log-ratio transformation of Shvayster & Peleg (1987) has been found to be an effective way to map $-\infty$ to ∞ into a range $\rho_{\min}$ to $\rho_{\max}$. Both of these transformations are contact transformations and so preserve the topological structure of the problem.

The second major problem in a general shift algorithm is the breakdown of conjugacy in the directions (Powell, 1977). The recursive formula used to generate a set of conjugate directions is prone to fail if the problem being minimized is not well conditioned, which can happen when some reflections are missing or much smaller than average, or if the search along each conjugate direction is not exact. It is necessary to monitor the conjugacy of the direction, by checking the angle between search directions, and reset when conjugacy breaks down.

Finally, there is the issue of the line search along each search direction. This could be done by Lagrange interpolation as with the pixel-based search, with the range of the search modified to echo the mapping function, but this is somewhat impractical as the search has to be able to include both very large and very small steps. It has been found effective to use a recursive search, first at a coarse interval and then at a fine one around the rough minimum. When a lookup table of error values already determined is used, this search averages about one Fourier transform per recursion (after the initial search) when the intervals are halved each time. This makes it possible to determine the best shift to arbitrary accuracy, typically 2^{-10}, without excessive cost.

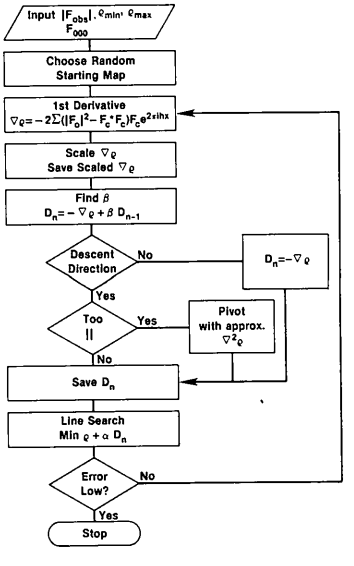

Fig. 1. The overall layout of a general direction conjugate gradient optimizer as applied to Patterson deconvolution.

The major costs of this algorithm are the evaluation of the gradient and the error. These both may be done with a FFT so that the cost per optimization step is $O(N \log N)$.

Choice of starting map

The starting map is chosen to be 'random'. This means that a pseudo-random number generator is used to generate an initial density. Pseudo-random number generators differ greatly in quality, and it is important to use a good one. The common linear congruential method (Knuth, 1981) should not be used because the short repeat length of these generators will make a highly structured starting map. In this work the Mitchel and Moore X_{n-24}, X_{55-n} additive random number generator was used (Knuth, 1981). It is also useful to apply a Dirichlet convolution followed by bounds constraint to the initial map so that the size of the Fourier coefficients for those reflections which are zero or unobserved is small. In addition, if space-group symmetry is used, it is important that the starting map be forced to have that symmetry.

Choice of search direction

The choice of the search direction is simple for observed reflections, the derivative of $\int dV\,(P_{\text{obs}} - P_{\text{calc}})^2$ with respect to F_{calc} described in the section on pseudo-inverses is the steepest descent direction and should be used. Unobserved data should be treated differently, and maximum-entropy or least-squares techniques can be used to fill the missing equations. If there is a small amount of missing data the data values can just be left to float. Otherwise it is important to minimize the 'power' or information in the unobserved data. This can be done by calculating gradients as though the unobserved data had a value of zero. A hybrid of allowing the data to float as long as its magnitude is sufficiently small, for example $|F_{\text{calc}}| \leq |\overline{F_{\text{obs}}}|$, is also useful. The error should only include the observed reflections. The Ψ function, defined below, is used to scale the derivative between -1 and 1 which prevents false convergence when the magnitude of the derivative is small. Space-group symmetry is maintained at this stage by requiring that the gradients have the correct symmetry in real space.

A more subtle numerical problem must also be treated in the choice of search directions. Since the

operations are defined on a grid, the effects of the Fourier transform of the grid must be minimized. Because of the sampling in real space, the calculated Fourier transform has errors caused by contamination from high-frequency terms not explicitly included in the summation. These effects are known informally as 'fold-over' and their effects on the FFT calculation of structure factors are well known (Ten Eyck, 1977; Sayre, 1951; Brillouin, 1962). 'Fold-over' effects can be strong enough to alter convergence even when the magnitudes are calculated from an image on the same size grid as the inversion. From multiple runs on varying grids, it appears that 'fold-over' error is a major cause of failure. The well known trick of forcing convergence in reciprocal space *via* an artificial thermal factor (Ten Eyck, 1977) is useful here. Since the gradients are calculated with the Fourier transform it is easy to add a small thermal factor to the gradients. If the high-resolution data are accurate, it is useful to make the thermal factor smaller as the error minimizes. The minimization is thus done in $\exp(-s^2)$ space, with little 'fold-over', but the error evaluation is still in a normal space. There are an infinite number of convergence factors in addition to the thermal factor (Young, 1988; Champeney, 1987; Phillips, 1984) and it is likely that a better choice can be found.

Enforcement of conjugacy

While almost any of the iterative formulae for enforcing conjugate directions (McCormick, 1983) are better than steepest descents, in general the Polak–Ribeire formula works best (Powell, 1977). In this procedure each search direction is the sum of the current steepest descent direction and a constant multiple of the last search direction.

$$d_0 = -\nabla F_0$$

$$d_{i+1} = -\nabla F_{i+1} + \beta d_i$$

$$\beta = \frac{\nabla F_{i+1} \nabla F_{i+1} - \nabla F_{i+1} \nabla F_i}{\nabla F_i \nabla F_i}$$

where d_i is the search direction and F_i is the function value.

This formula fails in two major ways (Powell, 1977). It can force the search direction to be a non-descent, and will produce a β which gives zero search directions when the gradients are nearly parallel. These conditions should be monitored and appropriate corrections taken. Since the dot product of the gradients between steps is found in order to find β, monitoring for parallel gradients is simple. When the gradients are nearly parallel for several steps in a row, an approximate second derivative is used to conjugate the shifts explicitly (Angel & Jain, 1978). This approximate conjugation can be used on its own, and is similar in convergence to the Polak–Ribiere form, but it is more expensive to calculate. The diagonal terms of the second derivative of the square error in the Patterson are just the calculated Patterson; this is circulant in form so the FFT can be used for efficient calculation.

$$d'_{i+1} = d_{i+1} - \beta d_i$$

$$\beta = d_{i+1} P d_i / d_i P d_i$$

where P is the Patterson in circulant form.

The angle between the search direction and the gradient is also monitored, and the search direction is reset to the gradient when this angle is larger than $\pi/4$.

Maintenance of constraints

There are three constraints which are used. The first is a lower bound on the density. The second is an upper bound. The third is a normalization constraint. With X-ray diffraction data the lower bound is zero; the upper bound depends on the image and for molecules five to ten times F_{000}/V_{cell} is an appropriate range.

In order to maintain the upper and lower bound constraints the log ratio mapping of Shvayster & Peleg (1987) is used. This can be broken into two functional steps, a scaling step and a mapping step. These are referred to as Ψ and Φ. The total mapping is $s = \Phi[\Psi(\rho)]$ with the inverse $\Psi^{-1}[\Phi^{-1}(s')]$, where s' is the modified output of the initial mapping. Note that without modifying the density $\Psi^{-1}(\Phi^{-1}\{\Phi[\Psi(\rho)]\})$ is just ρ. For a pair of bounds, $\rho_{\max}$ and $\rho_{\min}$, to the electron density ρ with ρ_{mean} as the mean value,

$$\Psi(\rho) = \frac{\rho - \rho_{\text{mean}}}{\rho_{\text{mean}}}$$

$$\Phi(\psi) = \log\left[\frac{\psi - \Psi(\rho_{\min})}{\Psi(\rho_{\max}) - \psi}\right]$$

$$\Phi^{-1}(s) = \left[\frac{\Psi(\rho_{\max})\, e^s + \Psi(\rho_{\min})}{1 + e^s}\right]$$

$$\Psi^{-1}(\varphi^{-1}) = \rho_{\text{mean}} \varphi^{-1} + \rho_{\text{mean}}.$$

When ρ_{mean} is zero, replacing it with $|\rho_{\min}|$ or $(\rho_{\max} - \rho_{\min})/2$ will work.

This mapping has a floating scale. If the modification of ρ wanders outside the bounds it is always remapped to the bounds. The scale between the map and the data will wander if it is not constrained. Without a scale constraint the process does not converge. This constraint is maintained by requiring that the output density have a constant F_{000}. This is done by multiplying the map by

$$F_{\text{obs}}(000)/F_{\text{calc}}(000).$$

Line search

Having defined a search direction and described a mapping which will keep the search in a bounded range of density values, we now describe the process of searching for a minimum error. Starting from a map ρ_0, the current best point, a map $\rho(D, \alpha)$ where D is the search direction and α is the length along D is generated. α is a scalar so that a one-dimensional search for the next minimum can be performed. Since the calculation of the error between the calculated and observed Pattersons is expensive, it is best to perform first a coarse search and then a fine search around the approximate minimum. If a table of error values as a function of α is kept and the search step is halved each time the cost of performing this search can be kept low.

The mapping functions Ψ and Φ, defined in the last section, are used in the generation of $\rho(D, \alpha)$. These are first applied to ρ_0, giving $s_0 = \Phi[\Psi(\rho_0)]$. s_0 has αD added to it, and the new map ρ is found by

$$\rho(D, \alpha) = \Phi^{-1}[\Psi^{-1}(s_0 + \alpha D)].$$

The Fourier transform of this map can then be used to find the error in the Patterson.

Convergence check

When the scaled square error in the Patterson

$$\frac{\int (P_{\text{obs}} - P_{\text{calc}})^2 \, \mathrm{d}V}{\int (P_{\text{obs}} - I_{000}/V_{\text{cell}})^2 \, \mathrm{d}V}$$

is less than about 10^{-4} the image has always been correct in model systems. This corresponds to a crystallographic R factor of about 0·001. Clearly such low numbers are not attainable in the presence of experimental errors. Therefore it is useful to consider convergence checks which are not dependent directly on the square error.

This is by no means a settled issue and the limitations of eigenvalue iteration discussed above apply to most convergence checks. One general scheme for a convergence check is to use a map improvement operator, and show that little or no change in the phases occurs. Unfortunately, the phase shift upon application of an operator to the density is not a reliable indicator of the accuracy of the phase set. If the solution and the operator are correct then little phase shift should occur, but it is not difficult to find incorrect phase sets which also give rise to limited phase shifts. In particular, the obvious map improvement operation of restoring the observed magnitudes to the structure factors, making the unobserved reflections zero, and applying positivity in real space is prone to form mixtures of enantiomorphs. This check should only be applied infrequently if there is any difficulty resolving enantiomorphs. With this caveat, the restoration of observed magnitudes, or even simply forcing unobserved data to zero, is an effective check.

Examples

In order to demonstrate that conjugate shifts are an effective algorithm, three examples will be given. The first of these is the reconstruction from intensities only of a small two-dimensional image. This shows that when the data are exact and complete the algorithm works. The second and third are the reconstruction from observed X-ray intensities of a 15-base-pair oligomer of DNA (Miller, Harrison, Appella, Wlodawer & Sussman, 1988) by the single-pixel algorithm and the general direction algorithm. This demonstrates that practical structure solution is possible in the presence of experimental error and incomplete data.

The duck

Fig. 2 shows the process of reconstructing a duck from its intensities. The image is on a 15 by 30 grid, with 450 degrees of freedom. Nearly all of the intensities are non-zero. An initial image was chosen with a random number generator, with the pixel values chosen to be between 0 and $2F_{000}/V_{\text{cell}}$. Each individual pixel was chosen at random, rather than choosing the phases at random and using the observed magnitudes followed by truncation. The scaled error converged from 0·994 to 3×10^{-13} in 49 steps, which is far less than the theoretical maximum of 450 steps. The total run time was about 5 min on a Silicon Graphics Personal IRIS. The image has a density range from 0 to 1, and the density limits were 0 and

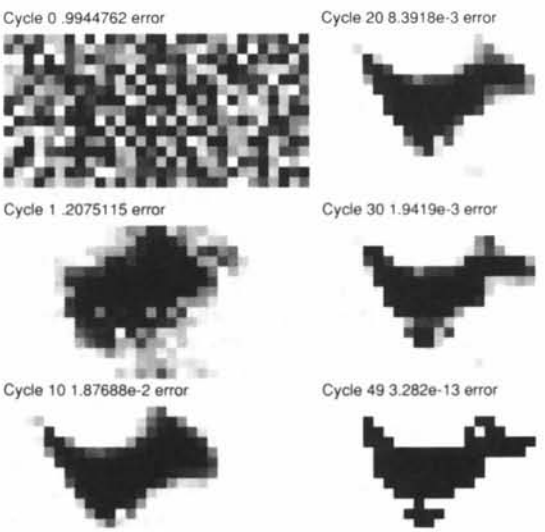

Fig. 2. The restoration of the image of a duck from its Fourier magnitudes only. The initial starting point is the output of a random number generator, and the final image is accurate to well within single-precision accuracy.

1·2. The exact value for the upper bound was not critical and convergence occurred when it ranged from 1 to about 1·8.

A real molecule

The crystal-structure solution of a 15-base-pair oligomer of DNA by minimum cross entropy and direct methods has been described elsewhere (Harrison, 1989; Miller *et al.*, 1988). The DNA crystallizes in space group $I222$ with a unit cell of $a = 36{\cdot}99$, $b = 53{\cdot}7$ and $c = 101{\cdot}6$ Å. The data used in the previous determination were about 75% complete to 3·0 Å and were used in these trials. The missing data are in the form of a wedge in reciprocal space which tests the ability of the algorithm to cope with systematic missing data. In both cases a grid of 30, 50, 100 was used; this is about 2/3 the Nyquist limit for 3 Å. I-centering zeros were treated as observed zeros. Figs. 3 and 4 show the results of both the single-pixel algorithm and the general search algorithm. These maps can be compared with Fig. 3 of Harrison (1989).

The single-pixel run for the DNA was performed in two passes on a Cray XMP 2-4. The initial model was a constant density of $3F_{000}/V_{\text{cell}}$. The first pass was six cycles to an error of 0·0361, followed by a run of minimum cross-entropy optimization and β iteration. The second was a further five cycles to an error of 4×10^{-5}. The F_{000} was 1537, which is about 1·5 times the true value. This and the absence of an upper bound allowed the error to get deceptively low. The initial pass required 1001 min c.p.u. time, and the second pass required 838 min for the total of 1839 min. There was little difference between the output map from the two passes. The 222 symmetry was enforced by operating with the $P222$ space-group-specific transform of a single pixel and optimizing over a single asymmetric unit. It is clear from the map that significant phase error remains, but the map is of better quality than that originally used to solve the structure.

The general direction run was done on a MIPS M/120, which is about 60 times slower than the Cray for vectorized code. From a random map with density values between 0 and $2F_{000}/V_{\text{cell}}$, the crystal symmetry and a Dirichlet convolution were applied, the results were truncated when less than zero and 548 cycles to an error of 0·0167 were run in 2384 c.p.u. min which would correspond to about 45 min on the Cray XMP 2-4. The program was run more than required for convergence, as the error was 0·019 after 100 cycles. The F_{000} was re-evaluated from the native Patterson to be 1004 so that the errors are more representative of the error in the map (if zero is used for I_{000} in the Patterson synthesis, then the lowest negative is approximately $-I_{000}/V_{\text{cell}}$). The lower-density bound was zero and the upper was 0·08, which corresponds to $12F_{000}/V_{\text{cell}}$. Zeros due to the I-centering were considered as observed zeros and counted in the error, while the unobserved data were simply restrained in magnitude. The $I222$ symmetry

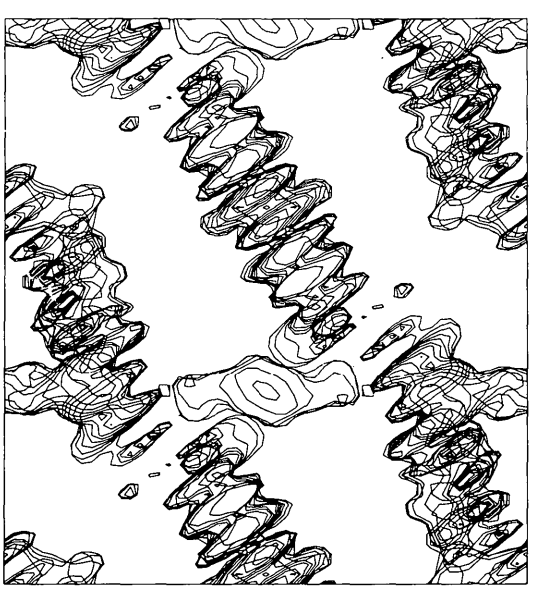

Fig. 3. A 20 Å slice through the electron density map of a 15mer of DNA found by the single-pixel algorithm. The map is one and a half times the a cell length on the vertical and one times the b cell length on the horizontal. The molecule runs diagonally from the upper center to the lower right corner.

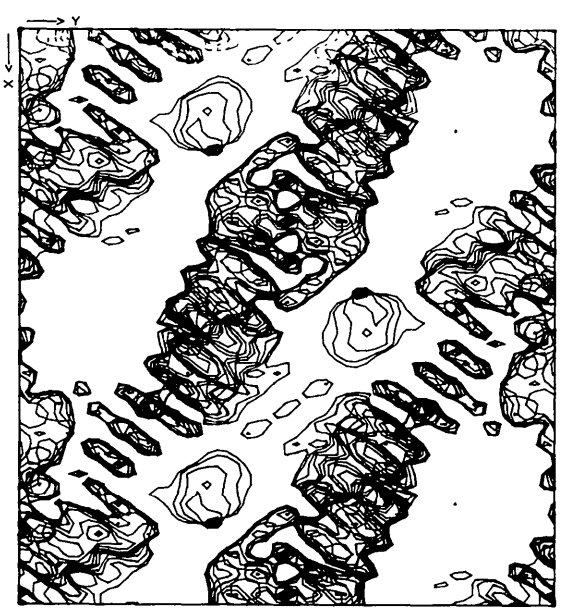

Fig. 4. The same 20 Å section as in Fig. 3. This density was found by the general direction search without the help of minimum cross-entropy processing. The molecule runs from the lower left corner to the upper center. Fewer centrosymmetric artifacts are visible than with the single pixel map. The two phase sets can be aligned on each other by origin shifts, but the maps have been left in their original reference frames to emphasize the arbitrariness of origin and enantiomorph when only the Fourier magnitudes are used.

was enforced in real space by performing the 222 symmetry operations on the gradient. The map shows less density accumulation on the symmetry axes than the single-pixel map and is of higher quality than the original map.

The map found by the general direction program (Fig. 4) is better in quality than the one found by the single-pixel search program (Fig. 3). The general direction program runs much faster, demonstrating the importance of algorithm design. While it would be possible to improve the performance of the single-pixel algorithm, at least in terms of map quality, on a serial or vector processor the relative cost weighs against it. However, both algorithms are considerably more effective than simulated annealing; tests on both pixel-based and general-direction-based annealing variables (Corana *et al.*, 1987) with two-dimensional test images suggest that the convergence of simulated annealing would be at least a thousand times slower than the pixel search.

Unresolved issues

The examples presented above demonstrate the potential of the algorithm. Several critical issues remain to be solved. The best way to choose F_{000}, $\rho_{\min}$ and $\rho_{\max}$ has not been found. Currently, the Patterson synthesis can be used for an estimate of F_{000}, but the best upper bound for the density must be found by trial and error. The effects of error in these choices is still not thoroughly understood. The manner in which space-group symmetry is enforced is sub-optimal. A major unresolved issue is the effect of different grids and data resolutions on convergence with real data. It is clear that 'fold-over' errors can prevent convergence. Various convergence functions (Young, 1988; Champeney, 1987; Phillips, 1984) exist, and combined with oversampling should reduce this effect. Finally, the issue of a convergence check is unsettled. Most convergence checks will fail for precisely the same reason that most density modification schemes will fail. Naive application of an operator like positivity, solvent flattening or squaring the map will fail to resolve mixtures of enantiomorphs. For the moment the best convergence check is to use several starting points and show that the solutions are similar. In addition, at this time, it is important to apply the programs to problems where molecular replacement could almost be used. By building confidence with essentially direct-method-assisted molecular replacement, it will be possible to develop the expertise needed to tackle completely new structures.

I would like to thank the Advanced Scientific Computing Laboratory, FCRF, for a substantial allocation of time on their Cray XMP and VAX 8600. The research was sponsored by the National Cancer Institute, DHHS, under contract No. N01-C0-74101 with BRI. The contents of this publication do not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the US Government.

References

ANGEL, E. S. & JAIN, A. K. (1978). *Appl. Opt.* **17**, 2186–2190.
BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410–445.
BRILLOUIN, L. (1962). *Science and Information Theory*, 2nd ed., pp. 93–97. New York: Academic Press.
CHAMPENEY, D. C. (1987). *A Handbook of Fourier Theorems.* Cambridge Univ. Press.
CONTE, S. D. & BOOR, C. DE (1980). *Elementary Numerical Analysis: an Algorithmic Approach*, 3rd ed. New York: McGraw-Hill.
CORANA, A., MARCHESI, M., MARTINI, C. & RIDELLA, S. (1987). *ACM Trans. Math. Software*, **13**, 262–280.
DAVIS, P. J. (1979). *Circulant Matrices.* New York: Wiley.
GONZALEZ, R. C. & WINTZ, P. (1977). *Digital Image Processing.* Reading, MA: Addison-Wesley.
HARRISON, R. W. (1989). *Acta Cryst.* A**45**, 4–10.
HAUPTMAN, H. A. (1972). *Crystal Structure Determination.* New York: Plenum.
HAYES, M. H. (1987). In *Image Recovery Theory and Application*, edited by H. STARK. Orlando, FL: Academic Press.
HENDRICKSON, W. A., SMITH, J. L. & SHERIFF, S. (1985). In *Methods in Enzymology: Diffraction Methods in Biology*, Vol. 115, part B, edited by H. W. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF. Orlando, FL: Academic Press.
ISREALEVITZ, D. & LIM, J. S. (1987). *IEEE Trans. Acoust. Speech Signal Process.* ASSP-**35**, 511–518.
KARLE, J. (1982). *Acta Cryst.* A**38**, 327–333.
KARLE, J. & HAUPTMAN, H. A. (1950). *Acta Cryst.* **3**, 181–187.
KNUTH, D. E. (1981). *The Art of Computer Programming*, Vol. 2. Reading, MA: Addison-Wesley.
LADD, M. F. C. & PALMER, R. A. (1980). In *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. LADD & R. A. PALMER. New York: Plenum.
LANE, R. G., FRIGHT, W. R. & BATES, R. H. T. (1987). *IEEE Trans. Acoust. Speech Signal Process.* ASSP-**35**, 520–525.
MCCORMICK, G. P. (1983). *Nonlinear Programming: Theory, Algorithms and Applications.* New York: Wiley.
MAIN, P. (1979). *Acta Cryst.* A**35**, 779–785.
MARVIN, D. A., BRYAN, R. K. & NAVE, C. (1987). *J. Mol. Biol.* **193**, 315–343.
MILLER, M., HARRISON, R., APPELLA, E., WLODAWER, A. & SUSSMAN, J. L. (1988). *Nature (London)*, **334**, 85–86.
PHILLIPS, E. R. (1984). *An Introduction to Analysis and Integration Theory.* New York: Dover.
POWELL, M. J. D. (1977). *Math. Program.* **12**, 241–254.
RAMACHANDRAN, G. N. & RAMAN, S. (1959). *Acta Cryst.* **12**, 957–964.
SAYRE, D. (1951). *Acta Cryst.* **4**, 362–367.
SHVAYSTER, H. & PELEG, S. (1987). *Pattern Recognition Lett.* **5**, 49–61.
TEN EYCK, L. F. (1977). *Acta Cryst.* A**33**, 486–492.
TULINSKY, A. (1985). In *Methods in Enzymology: Diffraction Methods in Biology*, Vol. 115, part B, edited by H. W. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF. Orlando, FL: Academic Press.
WANG, B. C. (1985). In *Methods in Enzymology: Diffraction Methods in Biology*, Vol. 115, part B, edited by H. W. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF. Orlando, FL: Academic Press.

WATENPAUGH, K. D. (1985). In *Methods in Enzymology: Diffraction Methods in Biology*, Vol. 115, part B, edited by H. W. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF. Orlando, FL: Academic Press.

WINOGRAD, S. (1978). *Math. Comput.* **32**, 141, 175-199.

YOUNG, N. (1988). *An Introduction to Hilbert Space.* Cambridge Univ. Press.

SHORT COMMUNICATIONS

Contributions intended for publication under this heading should be expressly so marked; they should not exceed about 1000 *words; they should be forwarded in the usual way to the appropriate Co-editor; they will be published as speedily as possible.*

Acta Cryst. (1990). A**46**, 619-620

Accurate computation of the rotation matrices.

By JORGE NAVAZA, *ER* 180 *du CNRS, Centre Pharmaceutique,* 92290 *Chatenay Malabry, France, and Immunologie Structurale, Institut Pasteur, Rue du Dr Roux,* 75015 *Paris, France*

(*Received* 22 *November* 1989; *accepted* 2 *January* 1990)

Abstract

A new recurrence relation between the reduced matrices of the irreducible representations of the rotation group is proposed, which permits their accurate computation for high orders of the representation.

This work was motivated by the appearance of numerical divergences during the computation of the fast rotation function (Crowther, 1972). The origin of this behaviour was found to be the numerical instability of the recurrence relation used to compute the rotation matrices, for moderately high angular momenta and a wide range of angles. Overflows were in fact detected for expansions involving spherical harmonics of order $j \geq 74$.

In an irreducible representation of the rotation group of dimension $2j+1$, the rotation parameterized by the Euler angles (α, β, γ) is represented by the matrix (Brink & Satchler, 1975)

$$D^j_{mn}(\alpha, \beta, \gamma) = d^j_{mn}(\beta) \exp -i(m\alpha + n\gamma). \quad (1)$$

The reduced matrices d^j_{mn} are determined by means of the 'triangular' relationship (Altmann & Bradley, 1963)

$$\begin{aligned}[(j-m)(j+m+1)]^{1/2} d^j_{m,n}(\beta) \\ +[(j-n+1)(j+n)]^{1/2} d^j_{m+1,n-1}(\beta) \\ +(m-n+1)\cot(\beta/2) d^j_{m+1,n}(\beta), \end{aligned} \quad (2)$$

starting from the analytical expression

$$\begin{aligned} d^j_{mj}(\beta) = (2j)!/[(j+m)!(j-m)!]^{1/2} \\ \times \sin(\beta/2)^{j-m} \cos(\beta/2)^{j+m}. \end{aligned} \quad (3)$$

Only the elements with $-n \leq m \leq n$, $n \geq 0$ have to be evaluated. According to (2) and (3), $(j-n+1)$, $(j-n+2)/2$ elements are necessary to determine d^j_{mn}. The number of operations grows as j^2 for small values of n, and the propagation of errors causes the observed divergences.

The unitarity of the representation implies the following orthogonality condition for the reduced matrices:

$$\sum_{n=-j}^{j} d^j_{m,n}(\beta) d^j_{m',n}(\beta) - \delta_{m,m'} = 0, \quad -j \leq m, m' \leq j. \quad (4)$$

Therefore, the magnitude of the errors produced by the numerical calculation may be described by computing the maximum, when m and m' are varied, of the absolute value of the left-hand member of (4), for given j and β. This is shown in Fig. 1, for $j \leq 60$, which are the values used in the standard program of Crowther. Although big enough, such errors do not produce overflows in most computers and are seldom detected.

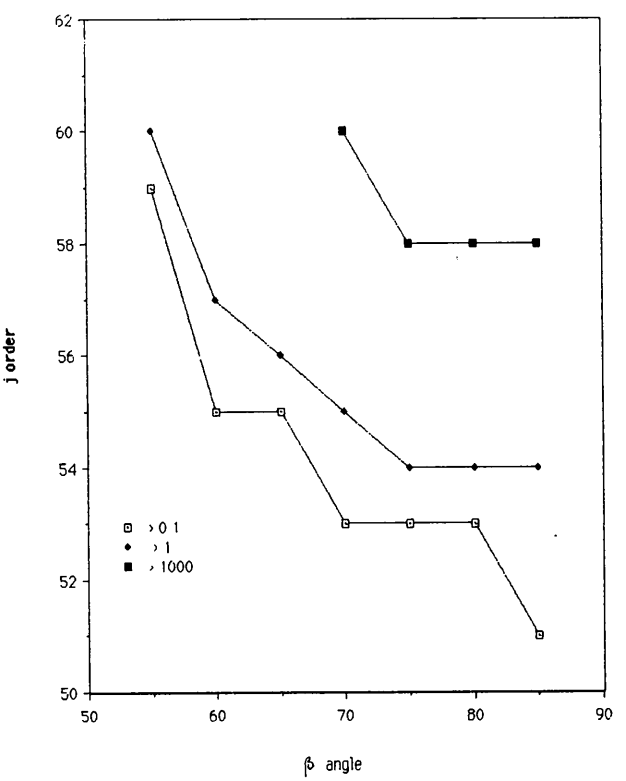

Fig. 1. Contour levels of the maximal deviation from the orthogonality conditions, of the reduced matrices $d^j(\beta)$, computed with the recurrence relation (2).

0108-7673/90/070619-02$03.00