leaves $\boldsymbol{\rho}_1'$ unchanged, in agreement with the requirement

$$\begin{pmatrix}\boldsymbol{\lambda}_1'\\ \sigma_1\end{pmatrix}=\mathbf{S}_2\boldsymbol{\rho}_3=\mathbf{S}_2^2\boldsymbol{\rho}_1=\begin{pmatrix}\mathbf{R}_2\boldsymbol{\lambda}_1\\ \sigma_1\end{pmatrix}. \quad (50)$$

Similarly, if the point marked ρ_1 in Fig. 3 now represents the direction of any position vector $\mathbf{d}$, then $\mathbf{d}$ may be augmented to homogeneous coordinates

$$\boldsymbol{\xi}=\begin{pmatrix}\mathbf{e}\\ f\end{pmatrix}=\begin{pmatrix}\mathbf{d}\\ 1\end{pmatrix}(1+d^2)^{-1/2} \quad (51)$$

and the normalized 4-vector $\boldsymbol{\xi}$ contains all the information in $\mathbf{d}$. Forming the product $\mathbf{R}_2\mathbf{d}$ is seen to be analogous to forming $\mathbf{S}_2^2\boldsymbol{\xi}$. In this sense the transformation of a position vector $\mathbf{d}$ and the compounding of two rotations are seen to be equivalent operations.

The vectors ±[1000], ±[0100] and ±[0010] represent 180° rotations about each of X, Y and Z, and ±[0001] gives the identity. It follows from (9) that the first three rows of $\mathbf{S}$, regarded as $\boldsymbol{\rho}$ vectors, correspond to the rotation $\boldsymbol{\rho}$ followed by 180° rotations about each of the reference axes, and the columns likewise correspond to $\boldsymbol{\rho}$ preceded by 180° rotations about them. Letting $\boldsymbol{\rho}_1$ be the first row of $\mathbf{S}$ and $\boldsymbol{\rho}_2$ the second column allows the corresponding ψ to be identified as the angle between the unrotated X axis and the Y axis rotated by $\boldsymbol{\rho}$, consistent with (45).

The top row of $\mathbf{S}$, as already stated, is the $\boldsymbol{\rho}$ vector which corresponds to the rotation $\boldsymbol{\rho}$ followed by a rotation of 180° about $\mathbf{X}$. This compound rotation has $\boldsymbol{\rho}$ vector $[-\sigma, \nu, -\mu, \lambda]$ and the corresponding $\mathbf{S}$ matrix is therefore

$$\begin{pmatrix}-\lambda & -\mu & -\nu & -\sigma\\ \mu & -\lambda & -\sigma & \nu\\ \nu & \sigma & -\lambda & -\mu\\ -\sigma & \nu & -\mu & \lambda\end{pmatrix}.$$

In the original rotation σ is algebraically distinct from λ, μ and ν in the ways in which it enters into the equations that arise. However, this example shows that the four rotations (the original, and three produced from it by compounding with 180° about X, Y and Z) collectively form a set in which λ, μ, ν and σ all have equivalent status and none is unique.

I am indebted to Drs A. D. McLachlan and R. Somorjai for helpful discussions.

References

AHARONOV, Y., FARACH, H. A. & POOLE, C. P. (1977). *Am. J. Phys.* **45**, 451-454.
COURANT, R. & HILBERT, D. (1953). *Methods of Mathematical Physics*, Vol. 1. New York: Wiley.
COURANT, R. & HILBERT, D. (1962). *Methods of Mathematical Physics*, Vol. 2. New York: Wiley.
DIAMOND, R. (1976). *Acta Cryst.* A**32**, 1-10.
KABSCH, W. (1976). *Acta Cryst.* A**32**, 922-923.
KABSCH, W. (1978). *Acta Cryst.* A**34**, 827-828.
LESK, A. M. (1986). *Acta Cryst.* A**42**, 110-113.
MCLACHLAN, A. D. (1972). *Acta Cryst.* A**28**, 656-657.
MCLACHLAN, A. D. (1979). *J. Mol. Biol.* **128**, 49-79.
MCLACHLAN, A. D. (1982). *Acta Cryst.* A**38**, 871-873.

Acta Cryst. (1988). A**44**, 216-222

Phase Extension by Combined Entropy Maximization and Solvent Flattening

BY E. PRINCE

Institute for Materials Science and Engineering, National Bureau of Standards, Gaithersburg, MD 20899, *USA*

AND L. SJÖLIN AND R. ALENLJUNG

*Institute for Inorganic Chemistry, Chalmers University of Technology and University of Göteborg, S-*41296 *Göteborg, Sweden*

(*Received* 8 *June* 1987; *accepted* 21 *October* 1987)

Abstract

An efficient algorithm is described for finding the maximum entropy density distribution under the constraint that $\langle\sum|F|\cos(2\pi\mathbf{h}\,.\,\mathbf{r}-\varphi)\rangle$, where the sum is over a subset of reflections whose phase has been determined, is constant. This algorithm is combined with solvent flattening in a procedure for extending phases to higher resolution. A test of the procedure on the structure of ribonuclease A and its application to the determination of two previously unknown structures are discussed.

Introduction

In crystallography, as in other branches of physics such as spectroscopy and radio astronomy, the observable data depend on Fourier transforms of density distributions that the experimenter wishes to determine. Because only the amplitude, not the phase,

0108-7673/88/020216-07$03.00

of the Fourier transform can be measured, and that only in a limited region of transform (reciprocal) space, and because the data are subject to random statistical fluctuations, the inverse transform that gives the density distribution is often poorly defined. In the study of biological macromolecules, various methods that have been described by the term 'density modification' have been used to try to improve the definition of the density map by incorporating prior knowledge in the computation of the inverse transform. The methods fall into several classes. 'Solvent flattening' (Wang, 1985) makes use of the fact that a protein is divided by a smooth surface into a protein region and a solvent region, in which the density is essentially constant at a known value. 'Wiener filters' (Wiener, 1949; Bode & Shannon, 1950; Robinson, 1967; Lacoss, 1971; Collins, 1978; Collins & Mahar, 1983*a*, *b*) make statistical predictions of the values of unobserved Fourier coefficients on the basis of the observed ones. A method that has become known as the method of maximum entropy uses information theory, as developed by Shannon (1948) and Jaynes [for a review see Jaynes (1979)] to obtain a distribution that is consistent with the observations, but is '"maximally noncommittal" with respect to all other matters'.

The maximum entropy method has been applied to the crystallographic problem by Collins (1982), by Wilkins, Varghese & Lehmann (1983) and by Bricogne (1984). Wilkins (1983) gives a numerical procedure for the solution of the maximum entropy problem by means of a 'single-pixel approximation'. This procedure, however, requires the iterative solution of a rather badly behaved transcendental equation for each pixel individually, which becomes cumbersome for macromolecular problems. This paper describes an efficient algorithm for the determination of a maximum entropy density distribution under the constraint that $\langle\sum|F|\cos(2\pi\mathbf{h}\,.\,\mathbf{r}-\varphi)\rangle$ is constant, where the sum is over a set of structure factors for which the phases have been previously determined. This algorithm is combined with solvent flattening to refine and extend phases to higher resolution. The procedure has been used successfully to reproduce the structure of ribonuclease A, starting from low-resolution phases, and to produce readily interpretable maps of the previously unknown structures of calcium-containing fragment 1 of bovine prothrombin (Sjölin, Alenljung, Svensson & Prince, 1988) and of fragment TR2C from bull testis calmodulin (Sjölin & Svensson, 1988).

Mathematical analysis

Consider a unit cell with volume V divided into N map elements, commonly referred to as 'pixels', with the mean density in the kth pixel denoted by ρ_k. The maximum-entropy approach consists in finding the maximum of

$$S = -\sum_{k=1}^{N} \rho_k \ln(\rho_k/q_k), \tag{1}$$

subject to the constraint that the sum of $(V/N)\rho_k$ over the unit cell must be equal to $F(000)$, and to one or more further constraints that require at least approximate agreement with the diffraction data. In (1) the quantities q_k represent a density distribution inferrred from prior information. Although various workers have considered how such prior information might be used to produce a nonuniform prior density (Bricogne, 1984; Collins, 1985), the analysis that follows assumes that the prior distribution is uniform, so that the terms in the sum reduce to $\rho_k \ln \rho_k$.

The most detailed analysis of the problem is due to Bricogne (1984), who suggests that the additional constraints be the individual fitting of many structure factors, with the phases of some and the amplitudes of the rest, with the maximum-entropy criterion being used to determine additional phases. He observes that there is a problem with phase extension, owing to the fact that there may be multiple sets of phases that have approximately the same entropy. Because this approach also becomes computationally unwieldy in macromolecular structures, Wilkins, Varghese & Lehmann (1983), following Gull & Daniell (1978), advocate using a 'weak' constraint of the form

$$\sum_{i=1}^{M} [(|F_{oj} - F_{cj}|)/\sigma_j]^2 \le M, \tag{2}$$

where the sum is over some subset of order M of the observed data, so that the distribution is statistically consistent with the observations. Previous workers have implemented this constraint by finding the minimum of a function of the form

$$Q(\boldsymbol{\rho}, \lambda) = \sum_{k=1}^{N} \rho_k \ln \rho_k + (\lambda/2) \sum_{j=1}^{M} [(|F_{oj} - F_{cj}|)/\sigma_j]^2, \tag{3}$$

where λ is commonly referred to as a Lagrange multiplier, although its actual use in practice seems to be as a relative weight between an entropy-maximization process and a least-squares process. [Note that λ appears as a constant factor multiplying the weights in the least-squares part of the expression. If λ is small, a perfectly uniform distribution will satisfy the inequality in (2), and the constraint does not apply. If λ is large the procedure reduces to a restrained least squares, with the entropy as the penalty function. Only if λ has exactly the value required to hold the sum of squares constant can it be considered to be a Lagrange multiplier. Although numerical methods exist for finding that value, they are computationally inefficient, and, as we shall see below, there are other approaches to the application of constraints that avoid the problem altogether.]

Constraints may be applied in a minimization (or maximization) process by partitioning parameter space into a constraint space spanned by (that is, reachable by a linear combination of) the constraint functions and a null space orthogonal to it. If a fitting algorithm starts at a point at which the constraints are satisfied, a *feasible point*, and makes moves only within the null space, the constraint conditions will remain satisfied. Alternatively, if the conditions for a stationary point in the null space can be expressed in an analytic form, a search can be made in the constraint space for a feasible point. An efficient means of making the partition is the variable-reduction method (Gill, Murray & Wright, 1981). To see how this method may be applied in the present context, consider an electron density map divided into N pixels, the densities in which are constrained to obey n constraint relations, $f_i(\boldsymbol{\rho}) = c_i$. Define an $n \times N$ constraint matrix, $\mathbf{C}$, by

$$C_{ik} = \partial f_i(\boldsymbol{\rho})/\partial \rho_k, \tag{4}$$

and partition $\mathbf{C}$ into an $n \times n$ square matrix, $\mathbf{V}$, and an $n \times (N-n)$ matrix, $\mathbf{U}$, so that $\mathbf{C} = (\mathbf{V}, \mathbf{U})$. Columns of $\mathbf{C}$ must be chosen so that $\mathbf{V}$ is non-singular. They are indicated here as the first n columns for convenience, but actually may be any set of n columns of $\mathbf{C}$. The $(N-n)$ rows of

$$\mathbf{Z} = [-(\mathbf{V}^{-1}\mathbf{U})^T, \mathbf{I}], \tag{5}$$

where $\mathbf{I}$ is the identity matrix of order $(N-n)$, are orthogonal to the rows of $\mathbf{C}$, and therefore form a basis set for the null space of the constraint relations. Note that the matrix $\mathbf{Z}$ is extremely sparse, having only $(n+1)\times(N-n)$ nonzero elements out of $N \times (N-n)$.

It is instructive to consider the solution by this method of a small hypothetical problem considered by Jaynes (1979). In this problem a standard die is cast many times. No record is kept of the number of times each face of the die appears on top, but only the long-term average number of spots per cast. If the die were fair, the expected value of the number of spots per cast would be 3·5, but in this case it is observed to be 4·5. The problem, then, is: what is the maximum-entropy distribution of probabilities, p_i, subject to the constraints

$$\sum_{i=1}^{6} p_i = 1 \tag{6a}$$

and

$$\sum_{i=1}^{6} i p_i = 4{\cdot}5? \tag{6b}$$

For this problem

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} \tag{7a}$$

and

$$\mathbf{Z} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 2 & -3 & 0 & 1 & 0 & 0 \\ 3 & -4 & 0 & 0 & 1 & 0 \\ 4 & -5 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{7b}$$

(Note that $Z_{j1} + Z_{j2} = -1$ for all j). For the maximum-entropy distribution of the p_k's we require a stationary point of $\sum p_k \ln p_k$ in the null space of $\mathbf{C}$. The rows of $\mathbf{Z}$ define a set of basis vectors for this null space, and the necessary and sufficient condition for a stationary point is that the partial derivatives of the sum with respect to moves parallel to each of these basis vectors vanish individually. Because $\partial(\sum p_j \ln p_j)/\partial p_k = \ln p_k + 1$, this condition is satisfied if

$$\ln p_k = (k-1)\ln p_2 - (k-2)\ln p_1 \quad \text{for } k = 3, 4, 5, 6. \tag{8}$$

Thus, the logarithm of any p_k is given as a linear function of two of them. Equation (8) may be rewritten

$$\ln p_k = \ln p_1 + (k-1)\ln(p_2/p_1), \tag{9a}$$

or

$$p_k = p_1 r^{(k-1)}, \tag{9b}$$

where $r = p_2/p_1$, so that, when the partial derivatives of the second constraint function with respect to the p_k's are in an arithmetic series, the p_k's themselves are in a geometric series. The sum of the p_k's must equal one, so that

$$p_1 = 1/\sum_{k=1}^{6} r^{(k-1)} = (r-1)/(r^6-1), \tag{10}$$

and the expected value of the number of spots is

$$\langle S \rangle = p_1 \sum_{k=1}^{6} k r^{(k-1)}, \tag{11}$$

which reduces to

$$\langle S \rangle = (1 - 7r^6 + 6r^7)/[(1-r)(1-r^6)], \tag{12}$$

in agreement with Jaynes's (1979) result.

$\langle S \rangle$ is a monotonically increasing function of r, the ratio of two of the probability densities, and the value of r corresponding to any value of $\langle S \rangle$ between one and six can be found by finding a root of a seventh-degree polynomial. In the macromolecular crystallography problem, of course, the number of pixels is of order 10^6 rather than six, but a number of principles seen in the die problem still apply. If only one quantity in addition to the overall normalization is constrained, the logarithms of the densities at every point for the maximum-entropy distribution are linear functions of the logarithms of the densities at only two points, and the normalization is a straightforward linear scaling. When the partial derivatives of the

second constraint function with respect to the pixel densities are on a linear scale, the corresponding densities are on an exponential scale, which tends to sharpen peaks and flatten valleys in the density distribution. Further, because $\exp(x)$ is positive for all x, the density distribution must necessarily be everywhere positive.

We shall now consider the question of a suitable constraint function to use for phase extension. Suppose the partial derivative of a constraint function with respect to the density in the kth pixel, ρ_k, is a linear function of the density, $\rho_k^{(0)}$, in a trial distribution before maximizing entropy,

$$\partial f(\boldsymbol{\rho})/\partial\rho_k = a + b\rho_k^{(0)}, \tag{13}$$

and suppose further that the pixels have been ordered in such a way that $\rho_1^{(0)} \neq \rho_2^{(0)}$. Then

$$\mathbf{V} = \begin{pmatrix} 1 & 1 \\ a+b\rho_1^{(0)} & a+b\rho_2^{(0)} \end{pmatrix} \tag{14}$$

is non-singular, and

$$\mathbf{V}^{-1} = \begin{pmatrix} (a+b\rho_2^{(0)})/\Delta & -1/\Delta \\ -(a+b\rho_1^{(0)})/\Delta & 1/\Delta \end{pmatrix}, \tag{15}$$

where $\Delta = V_{22} - V_{21} = b(\rho_2^{(0)} - \rho_1^{(0)})$. The first two elements of a typical row of $\mathbf{Z}$ are then $Z_{(k-2),1} = -(\rho_2^{(0)} - \rho_k^{(0)})/(\rho_2^{(0)} - \rho_1^{(0)})$, and $Z_{(k-2),2} = -(\rho_k^{(0)} - \rho_1^{(0)})/(\rho_2^{(0)} - \rho_1^{(0)})$, which are independent of both a and b, so that any such function will give the same $\mathbf{Z}$ matrix as the simplest one for which $a=0$ and $b=1$. Any function that is quadratic in ρ_k will meet the conditions for this to be true when evaluated at $\boldsymbol{\rho} = \boldsymbol{\rho}^{(0)}$, and, because $|F_c|$ is linear in ρ_k, many functions of $|F_c|^2$ fall in this class. Consider, for example, the function

$$E(\boldsymbol{\rho}) = \left\langle \sum_{k=1}^{M} |F_j| \cos(2\pi \mathbf{h}_j \,.\, \mathbf{r} - \varphi_j) \right\rangle, \tag{16}$$

the expected value for the density distribution $\boldsymbol{\rho}(\mathbf{r})$ of a Fourier density map computed from some subset of observed amplitudes and previously determined phases. Because the expected value of a sum is the sum of the expected values of its terms, and, if φ_j is the correct phase for F_j, $\langle \cos(2\pi \mathbf{h}_j \,.\, \mathbf{r} - \varphi_j) \rangle = |F_j|$, this reduces to

$$E(\boldsymbol{\rho}) = \sum_{j=1}^{M} |F_j|^2 \tag{17}$$

in the limit of correct phases. Now

$$\partial E(\boldsymbol{\rho})/\partial\rho_k = \sum_{j=1}^{M} |F_j| \cos(2\pi \mathbf{h}_j \,.\, \mathbf{r}_k - \varphi_j) = \rho_k^{(0)}, \tag{18}$$

so that entropy is maximized for constant $E(\boldsymbol{\rho})$ by $\boldsymbol{\rho}(\mathbf{r})$ such that

$$\ln\rho_k = [(\rho_2^{(0)} - \rho_k^{(0)}) \ln\rho_1^{(0)} + (\rho_k^{(0)} - \rho_1^{(0)}) \ln\rho_2]/(\rho_2^{(0)} - \rho_1^{(0)}), \tag{19}$$

where $\rho_2^{(0)}$ and $\rho_1^{(0)}$ are the densities in any pair of pixels such that $(\rho_2^{(0)} - \rho_1^{(0)}) \neq 0$.

Equation (19) gives a density distribution that maximizes entropy for constant $E(\boldsymbol{\rho})$ but not necessarily for $E(\boldsymbol{\rho}) = \sum |F_j|^2$. $E(\boldsymbol{\rho})$ is, however, a smooth monotonically increasing function of $x = (\ln\rho_2 - \ln\rho_1) = \ln(\rho_2/\rho_1)$. If we set $z_k = (\rho_k^{(0)} - \rho_1^{(0)})/(\rho_2^{(0)} - \rho_1^{(0)})$, equation (19) can be written

$$\ln\rho_k = \ln\rho_1 + z_k x, \tag{20a}$$

or, if we set the scale to give the correct value of $F(000)$,

$$\rho_k = [NF(000)/V] \exp(z_k x) / \sum_{j=1}^{N} \exp(z_j x). \tag{20b}$$

Now

$$E(\boldsymbol{\rho}) = \sum_{k=1}^{N} \rho_k^{(0)} \rho_k = [NF(000)/V] \sum_{k=1}^{N} \rho_k^{(0)} \times \exp(z_k x) / \sum_{k=1}^{N} \exp(z_k x), \tag{21}$$

and

$$\mathrm{d}E(\boldsymbol{\rho})/\mathrm{d}x = [NF(000)/V] \left\{ \sum_{k=1}^{N} z_k \rho_k^{(0)} \times \exp(z_k x) \Big/ \sum_{k=1}^{N} \exp(z_k x) - \left[\sum_{k=1}^{N} \rho_k^{(0)} \exp(z_k x) \right] \left[\sum_{k=1}^{N} z_k \exp(z_k x) \right] \times \left[\sum_{k=1}^{N} \exp(z_k x) \right]^{-2} \right\}. \tag{22}$$

With these relations, the equation $E(\boldsymbol{\rho}) = \sum |F|^2$ can easily be solved to any desired precision by the use of standard numerical methods. A distribution given by (19) has maxima, minima and saddle points in the same places as $\boldsymbol{\rho}^{(0)}(\mathbf{r})$, but the peaks are sharper, and the valleys are broader and flatter. The density must necessarily be everywhere positive, as is required for a maximum-entropy density. It should be noted that the representation of density as an exponential function is a basic requirement of Bricogne's (1984) maximum entropy equations. Furthermore, Collins & Mahar (1983*a*) have discussed this representation from a more practical viewpoint and have suggested that it can be a better representation of density than that given simply by a truncated Fourier series. Fig. 1 shows the result of the application of entropy maximization starting from $\boldsymbol{\rho}^{(0)}(x) = 1 + 2|F|\cos(2\pi x/a)$ for various values of $|F|$. The peak gets increasingly

sharp as $|F|$ gets larger, approaching a delta function as $|F|$ approaches 1·0.

On the assumption that a density map at slightly higher resolution will differ from this one only in detail, phases calculated from the Fourier transform of $\rho(\mathbf{r})$ may be used with the observed amplitudes to compute a new $\rho^{(0)}$ at somewhat higher resolution. Repeated application of this procedure at increasing resolution should ultimately lead to an everywhere-positive density distribution that is compatible with the amplitudes of all observed reflections. There is no guarantee, however, that it will not suffer from the lack of uniqueness identified by Bricogne (1984). That problem may be reduced by combining entropy maximization with molecular envelope definition and solvent flattening, as described by Wang (1985). A unified procedure then goes as follows: Start with a small set of phases, including three to define the origin, one to choose an enantiomorph, any available structure invariants and any others that have been determined by isomorphous replacement *etc.*, and with the observed amplitudes, and compute a low-resolution density map. Identify the molecular envelope and set the density in the solvent region constant. Within the envelope, find the pixels containing the minimum and maximum densities, and use the maximum density as ρ_2. If the minimum density is greater than a fraction, say 0·1, of the maximum density, use it as ρ_1. Otherwise, set ρ_1 arbitrarily equal to $0{\cdot}1\rho_2$. In either case, use the densities in those pixels as $\rho_2^{(0)}$ and $\rho_1^{(0)}$. This results in all values of z_k lying in the range from zero to one. Apply (19), and multiply the densities in all pixels by a constant factor to restore the overall normalization. The ratio ρ_2/ρ_1 is adjusted by application of the Newton-Raphson procedure to make $E(\rho)$ closer to the starting value, with the application of (19) and the renormalization then being repeated. Because the main objective is to determine phases that will lead to a non-negative density map at higher resolution, a rather crude solution to this one-dimensional fitting problem is acceptable, particularly in the early stages. Observation has shown, however, that the Newton-Raphson algorithm converges extremely rapidly. The modified density map is then Fourier transformed, and a new set of phases is computed, both for the reflections already used and some new ones. The whole procedure is then repeated until all reflections have been included, and all amplitudes and phases have converged to stable values.

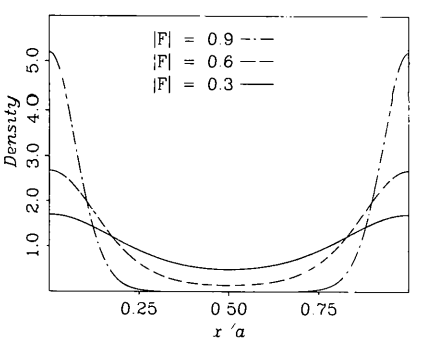


Fig. 1. Maximum entropy distributions, $\rho(x)$, corresponding to $\rho^{(0)}(x) = 1 + 2|F|\cos(2\pi x/a)$ for $|F| = 0{\cdot}3$, 0·6 and 0·9. The peaks get sharper and the valleys get flatter as $|F|$ increases.

Results

The maximum entropy procedure was coded as a subroutine in Fortran 77 in such a way that it could be easily incorporated in the commonly used density modification system described by Wang (1985). (Source code for this subroutine may be obtained at no charge from EP or LS.)

As a test of the procedure, we wanted to use a data set that contained phases to a nominal resolution of 2 Å or better. Ribonuclease A was chosen because its structure has been established to a high degree of precision (Wlodawer & Sjölin, 1983; Svensson, Sjölin, Gilliland, Finzel & Wlodawer, 1987; Wlodawer, Svensson, Sjölin & Gilliland, 1988). Fig. 2 is a schematic diagram of the procedure followed. In the first step the original phases were subjected to two cycles of density modification to get the maximum-entropy solution from what we consider to be the true structure. An appropriate envelope was then calculated. The same envelope was kept throughout the remainder of the investigation because our aim was to test how the procedure would work when the molecular object was known, and thus the position and the

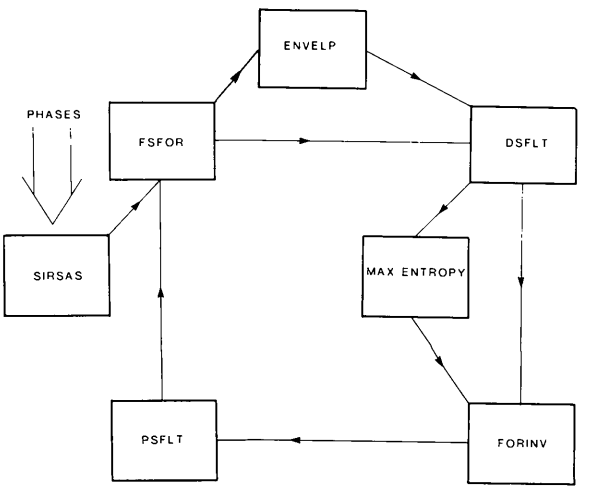


Fig. 2. A schematic diagram of the central program system according to Wang (1985) into which the maximum entropy routine has been inserted. *SIRSAS* loads phases into the general files; *FSFOR* performs direct-space Fourier synthesis; *ENVELP* calculates the molecular envelope; *DSFLT* performs direct-space filtering; *MAX ENTROPY* maximizes the entropy of the map; *FORINV* performs Fourier inversion; and *PSFLT* is a reciprocal-space filtering program.

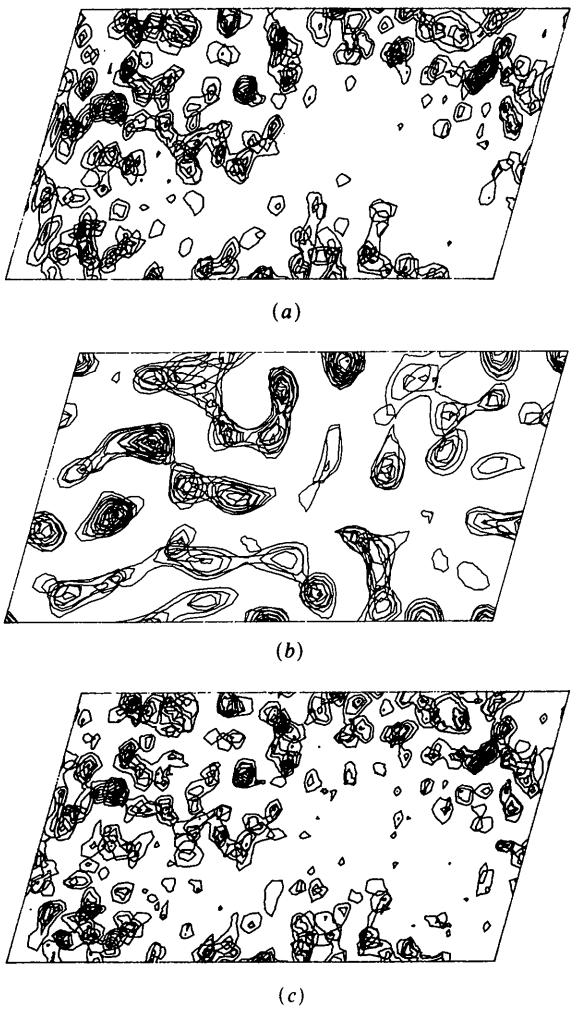
Table 1. *Summary of density modification according to Wang* (1985), *combined with maximum entropy calculations for phase extension in native Ribonuclease A*

Note that 'map type' contains information about the route in Fig. 2 that was chosen. F.o.m. indicates mean figure of merit.

Cycle number	Resolution	R value	F.o.m.	Phase shift	Map type
1	5·0	0·255	0·71	16·4	$F_o(DSFLT)$
2	5·0	0·207	0·73	19·5	$2F_o - F_m(MAX)$
3	4·5	0·224	0·72	14·1	$F_o(DSFLT)$
4	4·5	0·162	0·80	10·8	$2F_o - F_m(MAX)$
5	4·0	0·191	0·76	8·1	$F_o(DSFLT)$
6	4·0	0·140	0·83	7·2	$2F_o - F_m(MAX)$
7	3·5	0·177	0·78	6·0	$F_o(DSFLT)$
8	3·5	0·125	0·84	6·3	$2F_o - F_m(MAX)$
9	3·2	0·195	0·76	9·7	$F_o(DSFLT)$
10	3·2	0·141	0·82	9·0	$2F_o - F_m(MAX)$
11	2·7	0·172	0·78	6·1	$F_o(DSFLT)$
12	2·7	0·123	0·84	5·8	$2F_o - F_m(MAX)$
13	2·3	0·162	0·81	4·2	$F_o(DSFLT)$
14	2·3	0·128	0·83	5·2	$2F_o - F_m(MAX)$
15	2·0	0·153	0·83	3·5	$F_o(DSFLT)$

shape could be established to a high degree of precision. [The choice of envelope may prove to be the most important constraint in density modification (Schevitz, Podjarny, Zwick, Hughes & Sigler, 1981; Bhat & Blow, 1983; Podjarny, Bhat & Zwick, 1987). It remains to be seen how sensitive the phase extension procedure is to variations in the molecular boundary.] Two Fourier $F_o\alpha_c$ maps, using data to 2·0 and 5·0 Å resolution respectively, were calculated and contoured at a 1σ level. These contoured maps served as a reference with which later maps were compared. Phases to higher resolution than 5·0 Å were then stripped from the original data set and treated as unknown in the preliminary set up of the Wang density modification procedure. A series of calculations was then initiated according to the scheme in Fig. 2. During this series the resolution was extended in 15 steps; the parameters used and varied are summarized in Table 1. In each step an R value, defined as $\sum||F_o|-|F_m||/\sum|F_o|$, where F_m is the structure factor calculated by inversion of the modified map, and mean figure of merit as defined by Sim (1959) were calculated. (In principle, because the problem is underdetermined, it should be possible to reduce the R value to zero at each stage of phase extension. In practice, however, it appears to be unnecessary to iterate to full convergence to obtain a satisfactory set of new phases.) In alternate steps a $2F_o - F_m$ map was calculated instead of an F_o map in order to force F_m to become as close as possible to F_o. When the phases had been extended from 5·0 to 2·0 Å, a Fourier $F_o\alpha_c$ map was calculated and contoured. Fig. 3 shows a portion of each of three maps, one from the refined structure at 2·0 Å resolution, one with all reflections with resolution higher than 5·0 Å omitted, and the third with phases extended back to 2·0 Å resolution by the maximum-entropy procedure. As can be seen, the result from the phase extension procedure is in very good agreement with the original map. A calculation of the differences between the original phases and the extended phases revealed an average shift of roughly 35°.

The procedure was also tested on two unknown protein structures, fragment 1 from bovine prothrombin (see Fig. 4), in which the initial phases were obtained from multiple isomorphous replacement data, and fragment TR2C from bull testis calmodulin, where the phases were determined by the application of rotation functions to the published atomic coordinates for troponin-C (Herzberg & James, 1985). Each phase set was subjected to entropy maximization according to the scheme in Fig. 2, except that the molecular envelope was recalculated several times during the course of the calculations. In each case the quality of the map improved dramatically, and



(*a*)

(*b*)

(*c*)

Fig. 3. Overlaid sections of electron density maps from native ribonuclease A. (*a*) Map with phases to 2·0 Å resolution. (*b*) Map with phases to 5·0 Å. (*c*) Map with phases extended from 5·0 to 2·0 Å.

the structure determination could then proceed in a straightforward manner. More detailed descriptions of these applications will be published later (Sjölin, Alenljung, Svensson & Prince, 1988; Sjölin & Svensson, 1988).

The authors would like to thank D. M. Collins for a critical reading of the manuscript and many helpful suggestions.

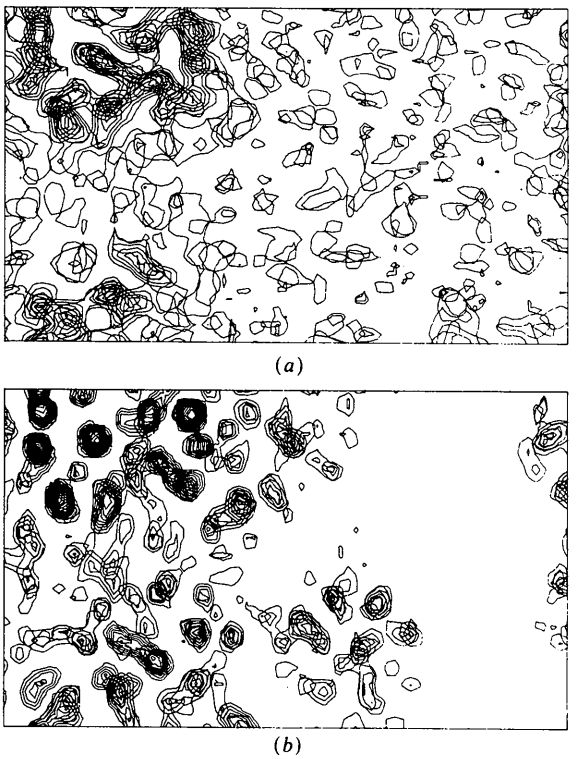


Fig. 4. Overlaid sections of electron density maps for calcium-containing fragment 1 of bovine prothrombin. (*a*) An F_o map with MIR phases to 3·2 Å. (*b*) Map with phases extended to 2·4 Å by solvent flattening and maximum entropy. Noise level in both maps set at 1σ level.

References

BHAT, T. N. & BLOW, D. M. (1983). *Acta Cryst.* A**39**, 166-170.
BODE, H. W. & SHANNON, C. E. (1950). *Proc. IRE*, **38**, 417-425.
BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410-445.
COLLINS, D. M. (1978). *Acta Cryst.* A**34**, 533-541.
COLLINS, D. M. (1982). *Nature (London)*, **298**, 49-51.
COLLINS, D. M. (1985). *Structure & Statistics in Crystallography*, edited by A. J. C. WILSON, pp. 109-112. Guilderland, NY: Adenine Press.
COLLINS, D. M. & MAHAR, M. C. (1983*a*). *Acta Cryst.* A**39**, 252-256.
COLLINS, D. M. & MAHAR, M. C. (1983*b*). *Acta Cryst.* A**39**, 777-783.
GILL, P. E., MURRAY, W. & WRIGHT, M. M. (1981). *Practical Optimization.* New York: Academic Press.
GULL, S. F. & DANIELL, G. J. (1978). *Nature (London)*, **272**, 686-690.
HERZBERG, O. & JAMES, M. N. G. (1985). *Nature (London)*, **313**, 653-659.
JAYNES, E. T. (1979). *The Maximum Entropy Formalism*, edited by R. D. LEVINE & M. TRIBUS, pp. 15-118. Cambridge, MA: Massachusetts Institute of Technology.
LACOSS, R. T. (1971). *Geophysics*, **36**, 661-675.
PODJARNY, A. D., BHAT, T. N. & ZWICK, M. (1987). *Ann. Rev. Biophys. Biophys. Chem.* **16**, 351-373.
ROBINSON, E. A. (1967). *Statistical Communication and Detection.* New York: Hafner.
SCHEVITZ, R. W., PODJARNY, A. D., ZWICK, M., HUGHES, J. J. & SIGLER, P. B. (1981). *Acta Cryst.* A**37**, 669-677.
SHANNON, C. E. (1948). *Bell Syst. Tech. J.* **27**, 379-423; 623-656.
SIM, G. A. (1959). *Acta Cryst.* **12**, 813-815.
SJÖLIN, L., ALENLJUNG, R., SVENSSON, L. A. & PRINCE, E. (1988). In preparation.
SJÖLIN, L. & SVENSSON, L. A. (1988). In preparation.
SVENSSON, L. A., SJÖLIN, L., GILLILAND, G. L., FINZEL, B. C. & WLODAWER, A. (1987). *Proteins: Structure Function Genetics.* In the press.
WANG, B. C. (1985). *Methods in Enzymology*, Vol. 114. *Diffraction Methods in Biological Macromolecules*, edited by H. W. WYCKOFF, C. H. W. HIRS & S. N. TIMASHEFF, pp. 114-167. New York: Academic Press.
WIENER, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* New York: John Wiley.
WILKINS, S. W. (1983). *Acta Cryst.* A**39**, 892-896.
WILKINS, S. W., VARGHESE, J. N. & LEHMANN, M. S. (1983). *Acta Cryst.* A**39**, 47-60.
WLODAWER, A. & SJÖLIN, L. (1983). *Biochemistry*, **22**, 2720-2728.
WLODAWER, A., SVENSSON, L. A., SJÖLIN, L. & GILLILAND, G. L. (1988). In preparation.

Acta Cryst. (1988). A**44**, 222-225

Some Thoughts on Harker–Kasper Inequalities

BY M. M. WOOLFSON

Department of Physics, University of York, York YO1 5DD, England

(*Received* 8 *August* 1987; *accepted* 27 *October* 1987)

Abstract

The approach by Harker & Kasper [*Acta Cryst.* (1948), **1**, 70-75] which led to the first inequality relationships between structure factors has not previously been applied to the space group $P1$ and there seems to have been a view that it could not give useful results for that space group. The idea has also been advanced that Harker-Kasper inequalities are contained within the complete set of determinantal

0108-7673/88/020222-04$03.00