Acta Cryst. (1993). D**49**, 148–157

Improvement of Macromolecular Electron-Density Maps by the Simultaneous Application of Real and Reciprocal Space Constraints

BY KEVIN D. COWTAN AND PETER MAIN

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

(*Received* 17 *May* 1992; *accepted* 29 *June* 1992)

Abstract

A general scheme for the improvement of electron-density maps is described which combines information from real and reciprocal space. The use of Sayre's equation, solvent flattening and histogram matching within this scheme has been described previously [Main (1990). *Acta Cryst.* A**46**, 372–377]. Non-crystallographic symmetry averaging, the use of a partial structure and constraints on individual structure factors have now been added. A computer program, *SQUASH*, is described which applies all these constraints simultaneously. Its application to the maps of several structures has been successful, particularly so when non-crystallographic symmetry is present. Uninterpretable maps have been improved to the point where a significant amount of the structure can be recognized. Applying the constraints simultaneously is more powerful than applying them all in series.

1. Introduction

Main (1990) has demonstrated how Sayre's equation (Sayre, 1952) and density-modification techniques may be combined in an effective and computationally efficient process for the phase refinement of protein structures. The solution of large non-linear systems such as Sayre's equation has previously presented computational problems. However, the use of the conjugate-gradient method to calculate shifts to the electron-density map and the use of fast Fourier transforms in performing convolution operations has rendered this calculation practical on many modern workstations. Zhang & Main (1990*b*) demonstrated the application of the procedure, known as *SQUASH*, to a small protein structure.

This paper describes recent progress with *SQUASH*. Equations are given for the constraint of individual structure factors in Sayre's equation and we describe additional density-modification techniques. Information from non-crystallographic symmetry and partial structures has now been combined with the original techniques of solvent flattening and density-histogram matching. Results are given for the application of *SQUASH* to a variety of known and unknown structures and we give a summary of our experiences in the use of the program.

2. Constraining the electron density

Let us consider a general method by which constraints in both real and reciprocal space can be combined to produce an estimate of the electron density. Typical real space constraints might include regions of known structure, or the results of density-modifiction calculations. Reciprocal space constraints include the measured structure-factor magnitudes and any known phase indications. Sayre's equation can be considered equally as a constraint in real or reciprocal space.

Each constraint is described in terms of a residual in real space which becomes zero when the constraint is satisfied. It may be written as

$$\mathcal{C}(\mathbf{x}) = 0 \tag{1}$$

where $\mathcal{C}$ is a function of the electron density $\rho(\mathbf{y})$ and $\mathbf{x}$ is the vector of parameters describing the constraint. Thus, we can form the residuals

$$r(\mathbf{x}) = \mathcal{C}(\mathbf{x}) \tag{2}$$

where the magnitude of the vector r is zero when $\rho(\mathbf{y})$ satisfies the constraint.

To find the value of the electron-density function ρ, we use the Newton–Raphson method, *i.e.* perform the iteration

$$\rho_{\text{new}} = \rho_{\text{old}} + \Delta\rho \tag{3}$$

where $\Delta\rho$ is given by the equation

$$J\Delta\rho = -r \tag{4}$$

and J is the Jacobian of the residual system

$$J(\mathbf{x},\mathbf{y}) = \partial r(\mathbf{x})/\partial\rho(\mathbf{y}). \tag{5}$$

The several constraints should be combined in such a way that the resulting electron-density function is most consistent with all of them. As the system of equations is overdetermined, this is obtained as the least-squares solution, which minimizes the combined residual of all the constraints. For M systems of equations, whose residuals and

0907-4449/93/010148-10$06.00

Jacobians are $r_i(\mathbf{x})$ and $J_i(\mathbf{x},\mathbf{y})$ respectively ($i = 1...M$), the normal equations of least squares take the form

$$\sum_{i=1}^{M} J_i^T J_i \Delta\rho = -\sum_{i=1}^{M} J_i^T r_i. \quad (6)$$

This is a linear system of N equations in N unknowns, where N is the number of grid points in the map. This system is solved for $\Delta\rho$ by the conjugate-gradient method, as described by Main (1990) [see also Sayre (1974)], which has the advantage that we neither need store the the large Jacobian J nor calculate the matrix J^TJ.

2.1 *Constraint systems*

We consider three types of constraint which can be applied to the electron density and their representation as real space residuals.

2.1.1. *Sayre's equation.* Sayre's equation is normally expressed in its reciprocal space form as

$$F(\mathbf{h}) = [\theta(h)/V]\sum_{k} F(\mathbf{k})F(\mathbf{h}-\mathbf{k}) \quad (7)$$

where $\theta(h) = f(h)/g(h)$ and $f(h)$ is the common atomic scattering factor, $g(h)$ the scattering factor of the squared atom. The real space residual is

$$r_0(\mathbf{x}) = (V/N)\sum_{y} \rho^2(\mathbf{y})\psi(\mathbf{x}-\mathbf{y}) - \rho(\mathbf{x}) \quad (8)$$

where $\psi(\mathbf{x})$ is the Fourier transform of $\theta(\mathbf{h})$. The equation may be interpreted as constraining the electron-density peaks to be the particular shape related to $\psi(\mathbf{x})$. It is applied here in the way described previously by Main (1990) and Zhang & Main (1990*b*).

2.1.2. *Density constraints.* There are several features of an electron-density map which may be known before the structure determination is complete and can be employed through the process of density modification. The techniques considered here are: solvent flattening, density-histogram matching, non-crystallographic symmetry averaging and partial structure information. Further comments on these techniques are given in the next section.

The combined techniques take an initial electron-density map and construct an improved map directly from it, as expressed by the equation

$$\rho(\mathbf{x}) = H(\mathbf{x}) \quad (9)$$

where $H(\mathbf{x})$ is the modified density.

We may want to apply different weights, or confidence values, to electron-density points. This allows the different density-modification techniques, which affect different areas of the map, to be weighted differently. Expressing the system of weights as $w_1(\mathbf{x})$, the residual equations are

$$r_1(\mathbf{x}) = w_1(\mathbf{x})[H(\mathbf{x}) - \rho(\mathbf{x})]. \quad (10)$$

2.1.3. *Structure-factor constraints.* We may wish to constrain the density to maintain consistency with the observed data. This is achieved by using structure factors which are known in both magnitude and phase, *e.g.* those specially well determined by multiple isomorphous replacement (MIR), and expressed in the equations

$$F(\mathbf{h}) = F_{\text{obs}}(\mathbf{h}) \quad (11)$$

where phases are included. Note that the $F(\mathbf{h})$ are variables and the $F_{\text{obs}}(\mathbf{h})$ are constants.

Applying weights $w_2(\mathbf{h})$ to individual structure factors, the reciprocal space residual takes the form

$$R_2(\mathbf{h}) = w_2(\mathbf{h})[F(\mathbf{h}) - F_{\text{obs}}(\mathbf{h})]. \quad (12)$$

The real space residual is the Fourier transform of this function where it takes the form of a convolution. Differentiation of the real space residual then gives the Jacobian for this system of equations.

This system has been applied successfully to constrain structure factors which have not yet been introduced into the calculation to zero. Unfortunately, attempts to constrain some structure factors to their MIR magnitude and phase, or by using a modified constraint equation to constrain structure factors in magnitude only, reduced the effectiveness of the phasing process. This unexpected result was found to be a consequence of the phase recombination (§4.5) in which figures of merit are calculated from the agreement between observed and calculated structure amplitudes. Constraining the magnitudes invalidates the figure-of-merit calculation.

2.2. *Applying the constraints*

The Jacobian for each constraint is calculated by differentiating the real space residual according to (5). The product of the Jacobian and its transpose with an arbitrary vector can then be formed. In each case, it simplifies to a series of convolutions which can be performed by fast Fourier transforms. The required expression for each constraint is listed in Table 1.

3. Density modification

As already described in §2.1.2, we calculate a modified map using whatever density information is available and constrain the electron density to equal the modified map.

The application of solvent flattening and histogram matching has been described by Zhang & Main (1990*a*) and the combination with Sayre's equation by Zhang & Main (1990*b*). Application of these techniques in the current work has not changed.

The new density-modification techniques in *SQUASH* are non-crystallographic symmetry aver-

Table 1. *Formulae for solving the compound system*

Constraint	J^TJp	J^Tr
Sayre's equation	$2\rho(\mathbf{x})\,\mathcal{F}^{-1}[\theta(-\mathbf{h})P'(\mathbf{h})] - \mathcal{F}^{-1}[P'(\mathbf{h})]$ where $P'(\mathbf{h}) = 2\theta(\mathbf{h})\,\mathcal{F}^{-1}[\rho(\mathbf{x})p(\mathbf{x})]$	$2\rho(\mathbf{x})\,\mathcal{F}^{-1}[\theta(-\mathbf{h})R'(\mathbf{h})] - \mathcal{F}^{-1}[R'(\mathbf{h})]$ where $R'(\mathbf{h}) = \theta(\mathbf{h})\,\mathcal{F}^{-1}[\rho^2(\mathbf{x})] - F(\mathbf{h})$
Density modification	$w_1^2(\mathbf{x})p(\mathbf{x})$	$w_1^2(\mathbf{x})[\rho(\mathbf{x}) - H(\mathbf{x})]$
Magnitude and phase	$\mathcal{F}^{-1}[P(\mathbf{h})w_2(\mathbf{h})w_2(-\mathbf{h})]$ where $P(\mathbf{x}) = \mathcal{F}^{-1}[p(\mathbf{x})]$	$\mathcal{F}^{-1}\{[F(\mathbf{h}) - F_{obs}(\mathbf{h})] \times w_2(\mathbf{h})w_2(-\mathbf{h})\}$

$\mathcal{F} \equiv (1/V)\sum_{\mathbf{x}}\exp(-2\pi i\mathbf{h}.\mathbf{x}/N)$ (discrete Fourier transform), $\mathcal{F}^{-1} \equiv (V/N)\sum_{\mathbf{h}}\exp(2\pi i\mathbf{h}.\mathbf{x}/N)$ (discrete inverse Fourier transform), p is an arbitrary vector, r is the residual vector for the particular constraint.

aging and the inclusion of partial structure information.

3.1. *Non-crystallographic symmetry*

Methods for the determination of the non-crystallographic symmetry elements are well established and it will be assumed here that the symmetry elements are known.

It is also necessary to determine an envelope describing which parts of the unit cell will be averaged together. For the tests described here, an approximation to the envelope was determined as follows. The geometrical centre of the non-crystallographic symmetry-related molecules was calculated from what was known about the structure and a Voronoi polyhedron constructed about it. This contained the region of the unit cell closer to the point than to any of its space-group symmetry equivalents. The envelope was then obtained from the overlap between this polyhedron and the protein mask determined by the method of Wang (1985).

The averaging calculation follows that of Bricogne (1974), except in the interpolation between grid points required in the transformation of the map by a non-crystallographic symmetry operator. It is usual to calculate an initial map at two or three times the data resolution and use linear interpolation to obtain density values from this map. In this work, it was convenient to calculate the map at the normal resolution in order to fit in with the other operations in *SQUASH*. Adequate results could be obtained using a map calculated at the resolution of the data and a quadratic interpolating function chosen to approximate the Fourier transform of the resolution sphere around the origin in real space. This interpolating function, like that of Bricogne, is convoluted with the map to give an estimate of the electron density at a non-grid location.

This method of interpolation was tested by comparing initial and transformed maps for a symmetrical density distribution. There is some loss of resolution in the interpolation process; however, the correlation coefficient between the intial and transformed maps was found to be greater than 0.96 over a wide range of tests. The transformed map may be rescaled to keep its variance the same as that of the initial map.

Averaging is performed between the initial and transformed maps within the chosen envelope and the unit cell reconstructed from the averaged subunit. The non-crystallographic symmetry averaging calculation can be applied either on its own or together with histogram matching and solvent flattening. In the latter case, it is best to apply the symmetry averaging first, as the loss of resolution in the interpolation process tends to corrupt the electron-density histogram.

3.2. *Partial structure*

It is common at some stage of the structure determination to have an atomic model of parts of the molecule. Such partial structure information is usually combined with experimental data in reciprocal space, using the method of Read (1986) or similar. This method takes into acount the effect of

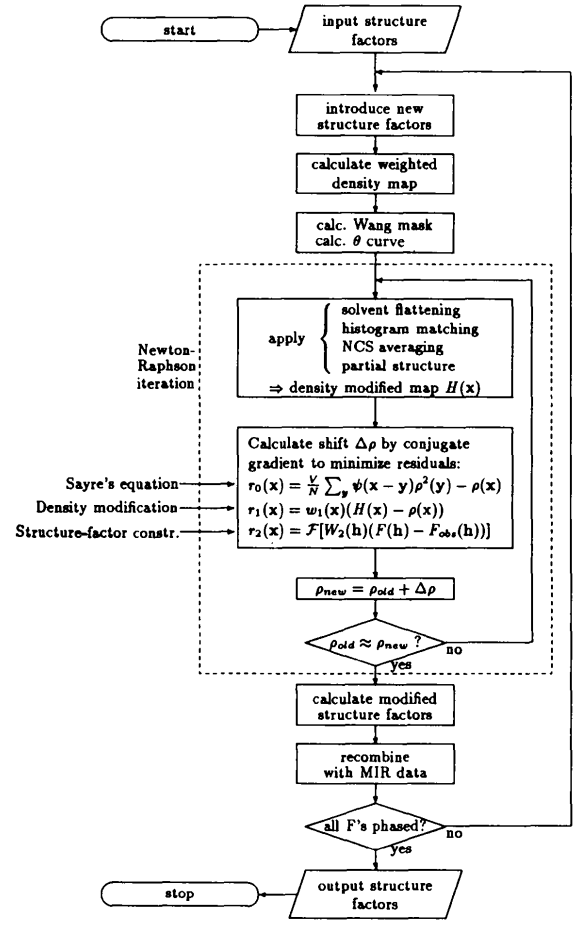


Fig. 1. Flowchart for phase refinement by Sayre's equation and density modification.

errors in the data upon each reflection and is superior to density-modification methods for normal use. However, if we are already applying density modification, it is useful to include the partial structure in the density-modification process as well.

This is achieved by calculating an electron-density map from the model coordinates and then averaging this with the estimated density in the regions where the model is available. To avoid biasing the map, the coordinate errors in the model, estimated by the method of Read, must be taken into account.

The effect of errors in the atomic coordinates can be simulated by convoluting the atomic density with a Gaussian function. This is equivalent to multiplying the partial structure factors by a temperature factor in reciprocal space. The effective temperature factor which must be applied (Blundell & Johnson, 1976) is given by

$$B_{\rm eff} = 8\pi^2\overline{u^2} \qquad (13)$$

where $\overline{u^2}$ is the mean-square error in atomic position. In addition, if the mean coordinate error is non-zero, it is necessary to take a weighted average of the initial and model densities

$$\rho_{\rm mod} = (1 - w_{\rm par})\rho_{\rm initial} + w_{\rm par}\rho_{\rm par}. \qquad (14)$$

The weighting function should be unity for no errors and zero for infinite errors. A satisfactory empirical function is

$$w_{\rm par} = \overline{\exp(-B_{\rm eff}\sin^2\theta/\lambda^2)} \qquad (15)$$

where the average is taken over all reflections.

4. Practical considerations

4.1. *Program outline*

A program, based on that of Zhang (1989), was constructed to test Sayre's equation and the constraint systems on protein structures. The program assumes that structure-factor magnitudes and phases with figures of merit are available at some starting resolution. It is possible either to refine phases at that resolution or to determine phases for magnitudes at higher resolution if they are available. In addition, structure factors may be calculated in both magnitude and phase if they are missing from the original data.

A flow chart of the calculation is given in Fig. 1. Note that the density modification is performed inside the Newton–Raphson loop. The density-modified map is recalculated after each Newton–Raphson iteration on the basis of the improved map obtained from the previous cycle. The right-hand side of the density-modification equations (9) therefore changes between iterations, making this apparently linear system non-linear in a complicated and unpredictable manner.

4.2. *Scaling the input data*

It should be noted that in several places, electron density or stucture factors from different sources are compared and combined. It is therefore of vital importance, especially in the phase recombination (§4.5) that the input data be on an absolute scale. The Wilson plot (Wilson, 1949) was not sufficiently accurate for this purpose, so the following, more robust, method was used instead.

The electron-density distribution and solvent level fix the mean and variance of the whole electron-density map. Thus we can scale the input data to be consistent with the target histogram using the following relationships, obtained from the structure-factor equation and Parseval's theorem:

$$\bar{\rho} = (1/V)F(000) \qquad (16)$$

$$\sigma_\rho^2 = (1/V^2)\sum_{h}|F(\mathbf{h})|^2 \qquad (17)$$

where σ_ρ^2 is the variance of the electron density about zero.

The mean and variance of the electron-density map at the desired resolution is calculated using the target histogram, the mean value of the solvent density and the solvent content of the cell. $F(000)$ can then be evaluated from (16) and the scale of the input magnitudes from (17). This method is adequate for scaling unknown data sets at any resolution.

4.3. *The θ curve*

To apply Sayre's equation, we must first calculate the function $\theta(h)$, where $h = |\mathbf{h}| = 2\sin\theta/\lambda$ [equation (7)]. At infinite resolution, we expect θ to be a spherically symmetric function which decreases smoothly with increasing $\sin\theta$. However, for data at less than atomic resolution, the θ curve will behave differently because atomic overlap changes the peak shapes. In practice, it is calculated empirically from Sayre's equation itself using the formula

$$\theta(h) = V\left\langle \frac{F(\mathbf{h})}{\sum F(\mathbf{k})F(\mathbf{h}-\mathbf{k})} \right\rangle_h \qquad (18)$$

where the average is carried out over ranges of $|\mathbf{h}|$, *i.e.* over spherical shells each covering a narrow resolution range.

For phase extension, a Gaussian function of $|\mathbf{h}|$ is fitted to the available values of θ and used as an estimate of θ in the high-resolution region. The use of a weighted combination of the empirical and Gaussian curves, based on the number of reflections in the range, avoids problems when there are too few reflections to obtain a reliable average.

θ curves produced by this method for insulin are shown in Fig. 2. They are plotted as a function of $\sin^2\theta/\lambda^2$ at the starting resolution of 3.0 Å and after phase extension to 2.15 Å, respectively. Note that the

final curve in the extension region closely matches the original extrapolated curve. Various formulae based on analytical approximations to the atomic shape have also been used to estimate θ. However, the empirical method given here has been shown to be the most reliable over a range of structures.

4.4. *Phase extension*

It is normal in density-modification calculations to extend the resolution in small steps (Hendrickson & Lattman, 1970; Zhang, 1989), the resolution being increased slightly after each phase recombination. This is in contrast to small-molecule direct methods, where it is normal to phase the largest structure factors first as these have the greatest affect on the density map. Our own results were improved by introducing reflections on the basis of a weighted combination of both the reflection resolution and its magnitude. The weight function used is

$$w_{\text{intro}} = |F(\mathbf{h})|^2\left[1 - \tfrac{3}{4}\left(\frac{s}{s_{\max}}\right) + \tfrac{1}{16}\left(\frac{s}{s_{\max}}\right)^3\right] \quad (19)$$

where $s = |\mathbf{h}|$ and $s_{\max}$ is the value of s at the resolution limit. The term in square brackets is proportional to the number of terms available in Sayre's equation for phasing that particular reflection. At each stage of the refinement, all the reflections with a weight above a certain threshold value are included. Those left out of the calculation are set to zero at the beginning of each cycle of phase extension. The threshold value is reduced between cycles of phase extension until all structure factors have been included.

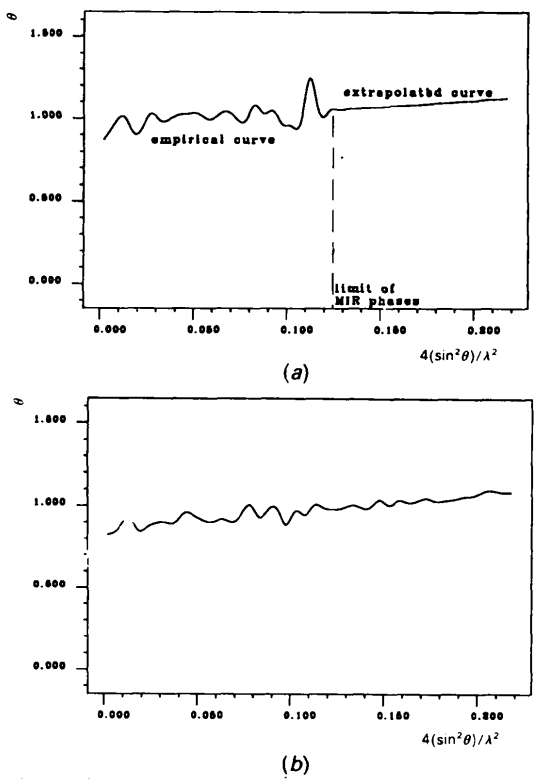


Fig. 2. (*a*) The empirical θ curve calculated from Sayre's equation using 3.0 Å MIR data. Extrapolation is achieved by fitting a Gaussian to the empirical curve. (*b*) The θ curve calculated from Sayre's equation after phase extension to 2.15 Å.

4.5. *Phase recombination*

Once the electron density has been modified to make it consistent with the available information, the map is transformed to give a new set of structure factors, modified in both magnitude and phase. These are then combined with the experimental data. It is normally assumed that the F_{obs} are accurate and that the phases determined from experimental methods have a certain probability distribution about the estimated values.

Figures of merit for the modified structure factors are estimated from the size of the observed and modified structure-factor magnitudes by the formula

$$\text{FOM} = I_1(X)/I_0(X) \quad (20)$$

where I_0 and I_1 are zero- and first-order modified Bessel functions and

$$X = \frac{2\sigma_A|E_{\text{obs}}||E_{\text{mod}}|}{1 - \sigma_A^2} \quad (21)$$

where E_{obs}, E_{mod} are normalized observed and modified structure factors and σ_A is as measure of the accuracy of the modified map (Srinivasan, 1966). The parameter σ_A is estimated by the method of Read (1986).

Once a figure of merit has been estimated for the modified phase, it can be combined with the MIR phase by the method of Hendrickson & Lattman (1970). Thus we arrive at a combined phase in which the information from the density-modification calculation has been filtered by the agreement between the modified magnitudes and the MIR data.

Read's method for estimating the parameter X is found to give better results than that of Sim (1959), provided there are sufficient reflections to give reliable statistics.

4.6. *Computational requirement*

Most of our calculations have been performed on Silicon Graphics MIPS-3000 based machines. The CPU time for a phase extension using Sayre's equation and density modification from 3 to 2 Å on a small protein like insulin is about 1 h. Much of this is in the calculation of fast Fourier transforms, so the time increases slightly faster than the unit-cell volume.

Table 2. *The test structures*

Structure	Full name	Unit cell	Space group	No. of residues	Solvent content (%)	NCS*	MIR data resolution (Å)	Native data resolution (Å)	References
Insulin	2Zn pig insulin	$a = 82.5, b = 82.5, c = 34.0$ Å $\alpha = 90, \beta = 90, \gamma = 120°$	$R3$	51	30	Twofold rotation	3.0	2.1	Adams *et al.* (1969), Baker *et al.* (1988)
RNase	Guanyloribonuclease from *Streptomyces aureofaciens*	$a = 64.9, b = 78.3, c = 38.8$ Å $\alpha = 90, \beta = 90, \gamma = 90°$	$P2_12_12_1$	96	45	—	3.0	2.5	Ševčík *et al.* (1981), Ševčík & Zelinka (1984)
IL1β	Recombinant Human interleukin 1β	$a = 55.0, \beta = 55.0, c = 77.1$ Å $\alpha = 90, \beta = 90, \gamma = 90°$	$P4_3$	153	70		3.2	2.2	Oldfield *et al.* (1992), Finzel *et al.* (1989)
CHMI	5-Carboxymethyl-2-hydroxy-muconate isomerase	$a = 90.4, b = 90.4, c = 130.1$ Å $\alpha = 90, \beta = 90, \gamma = 90°$	$P4_12_12$	125	52	Threefold rotation	3.7	2.1	Wigley *et al.* (1992)
P-A	Penicillin acylase	$a = 51.9, b = 64.6, c = 75.73$ Å $\alpha = 100.1, \beta = 111.4, \gamma = 106.1°$	$P1$	766	45	—	5.0–3.5	2.5	Duggleby *et al.* (1992)

* Non-crystallographic symmetry.

The major limitation is the amount of memory required. The conjugate-gradient calculation requires at least six maps in memory at once and, in practice, it is useful to be able to hold more than this. For insulin, about 12 MB of real memory are required and this figure increases in proportion to the unit-cell volume. Fewer maps are required if diagonal approximation is used and this also reduces the computing time by a factor of about 5. A small further reduction of both time and memory is achieved if Sayre's equation and the reciprocal space constraint are excluded from the calculation.

5. Results

Details of five structures to which we have applied *SQUASH* are given in Table 2. Two of the structures, insulin and RNase, were known when work was started on them. Interleukin 1β and 5-carboxymethyl-2-hydroxymuconate isomerase became known while we were testing *SQUASH* on them and were useful in its development. The final structure, penicillin acylase, is currently being determined with the aid of *SQUASH*.

To indicate the effectiveness of the different techniques described in this paper, they were applied in various combinations to the four known test structures. The following measures of the quality of the results were calculated:

(*a*) Unweighted mean phase error between *SQUASH* phases and those of the known structure.

(*b*) The correlation coefficient between density maps obtained from *SQUASH* and from the known structure. The correlation coefficient is calculated from

$$\text{correlation} = \frac{\overline{\rho_1\rho_2} - \overline{\rho_1}\,\overline{\rho_2}}{[(\overline{\rho_1^2} - \overline{\rho_1}^2)(\overline{\rho_2^2} - \overline{\rho_2}^2)]^{1/2}} \qquad (22)$$

where ρ_1 and ρ_2 are the density maps to be compared. This comparison is performed both for a map using *SQUASH* phases only and for a weighted map using the figures of merit for the weights. This gives an indication of how good the figures of merit are.

Table 3. *Phase refinement results*

Data set	Mean phase error (°) Initial	Extended	All	Correlation φ	Correlation φ and FOM
Insulin					
MIR	46.9	—	—	0.401	0.555
HM/SF	42.4	73.3	61.4	0.584	0.626
SAYR/HM/SF	40.9	68.2	57.7	0.639	0.671
NCS	44.4	92.1	73.8	0.419	0.547
HM/SF/NCS	38.6	70.2	58.1	0.640	0.681
SAYR/HM/SF/NCS	38.4	67.7	56.5	0.658	0.697
RNase					
MIR	59.7	—	—	0.300	0.396
HM/SF	55.8	80.3	65.9	0.429	0.470
SAYR/HM/SF	55.3	78.5	64.9	0.443	0.484
IL1β					
MIR	67.5	—	—	0.179	0.210
HM/SF	70.4	84.3	78.6	0.222	0.262
SAYR/HM/SF	70.5	86.5	80.0	0.195	0.244
CHMI					
MIR	72.9	88.6	85.5	0.086	0.139
HM/SF	70.3	85.8	82.1	0.145	0.184
SAYR/HM/SF	70.3	85.4	82.4	0.149	0.185
NCS	70.3	88.7	85.0	0.107	0.150
HM/SF/NCS	64.5	73.9	72.0	0.328	0.394
CHMI partial structure data					
COMB	69.7	81.9	77.5	0.245	0.271
COMB + HM/SF/PAR	68.6	80.5	76.2	0.265	0.293

Abbreviations: MIR, initial phases input to *SQUASH*; COMB, combined MIR and partial structure phases; SAYR, Sayre's equation; SF, solvent flattening; HM, histogram matching; NCS, non-crystallographic symmetry averaging; PAR, partial structure.

Comparisons between the estimated and calculated phases for the initial MIR data and for the *SQUASH* results are listed in Table 3. These are discussed in the following sections.

5.1. *Insulin*

The primitive cell of insulin contains six molecules, with pairs forming dimers related by a non-crystallographic twofold axis and the dimers related by a crystallographic threefold axis. The structure was originally solved directly from a 3 Å MIR map (Adams *et al.*, 1969).

Table 3 shows that histogram matching (HM) and solvent flattening (SF) together are effective both for phase refinement and phase extrapolation. Inclusion

of Sayre's equation (SAYR) leads to a further improvement. It is evident from the mean phase error of the extended phases that Sayre's equation is more powerful than density modification for phase extension at this resolution.

In applying non-crystallographic symmetry (NCS), the phases were extrapolated in the same way as with other density-modification techniques. This is different from normal practice where the extension is usually carried out very slowly and carefully. Failure to do this results in the extrapolated phases containing little useful information as can be seen from Table 3. However, the combination of non-crystallographic symmetry averaging with histogram matching and solvent flattening leads to a great improvement in both initial and extrapolated phases. Histogram matching and solvent flattening are effective in counteracting the loss of resolution in the averaging calculation and in increasing the resolution of the map. The inclusion of Sayre's equation gives rise to a further improvement.

5.2. *RNase*

RNase is similar to insulin both in size of cell and quality of data. MIR phases are available to 2.5 Å, which is also the limit of the observable data. The MIR phases and figures of merit were removed for all structure factors beyond 3 Å resolution and then phase extension was performed to 2.5 Å.

The results in Table 3 show that the density modification made a significant improvement to the map. However, the phase extension is less satisfactory than for insulin. In particular, Sayre's equation was less effective. This is a result of the poorer quality of the starting map and the lower resolution of the final map.

5.3. *Interleukin 1β*

The experimental data of interleukin was phased to 3.2 Å, with native magnitudes to 2.2 Å resolution. The MIR map was uninterpretable. It was even very difficult to distinguish the solvent from the protein regions of the map.

This was a much worse initial map than in previous tests, but density modification still gives some improvement. The final map shows solvent and protein regions fairly clearly and it contains some interpretable features. The addition of Sayre's equation gives poorer results because of the large volume of solvent and the large amount of noise in the initial map.

5.4. *5-Carboxymethyl-2-hydroxymuconate isomerase*

The experimental data consists of native magnitudes to 2.1 Å resolution and phase information to 3.7 Å. One derivative provided weak phase information to 2.6 Å. The initial MIR map was uninterpretable. Phase extension was performed from 3.7 to 2.1 Å, although the phase information in the extension region was also taken into account in the phase recombination. It can be seen from Table 3 that the effects of solvent flattening, histogram matching and Sayre's equation are similar to those found previously.

The application of the threefold non-crystallographic symmetry on its own makes only a slight improvement in the map although, as before, there is no significant phase extension. However, the combination of non-crystallographic symmetry averaging, solvent flattening and histogram matching leads to a much greater improvement. This map is significantly better than the one from which the structure was initially determined and there is little doubt that much of it can be interpreted.

Fig. 3 compares sections of the map calculated from MIR phases and *SQUASH* phases. The contours are at intervals of 1.5 standard deviations for each map. Correct atomic positions are superimposed for $x < y$ only. The MIR map shows few interpretable features and even isolating the protein from the solvent regions is difficult. In the density-modifed map, however, there is clear contrast between solvent and protein regions and these correspond closely with the actual molecular position. The density peaks also agree well with the correct atomic positions.

5.4.1. *Partial structure.* The initial atomic model, containing 30% of the main chain with an estimated r.m.s. coordinate error of 0.8 Å was available for this structure (Wigley *et al.*, 1992). This provided a realistic test of the use of partial structure information in the density modification. The MIR phases and partial structure data were combined using Read's σ_A method to give a starting map. The statistics for this map, before and after the application of density modification, are shown in Table 3.

The combined phases, calculated from the MIR and model phases, are considerably better than the MIR phases. Density modification, which includes the partial structure, gives a further improvement. Examination of the maps reveals that the partial structure affects the region of the cell local to the model and has little effect elsewhere.

5.5. *Penicillin acylase*

The initial data set consists of native magnitudes to 2.5 Å resolution and MIR phases to 3.5 Å although, owing to lack of isomorphism, the phase information beyond 5.0 Å is very weak.

The initial MIR map was uninterpretable. Histogram matching and solvent flattening were applied to

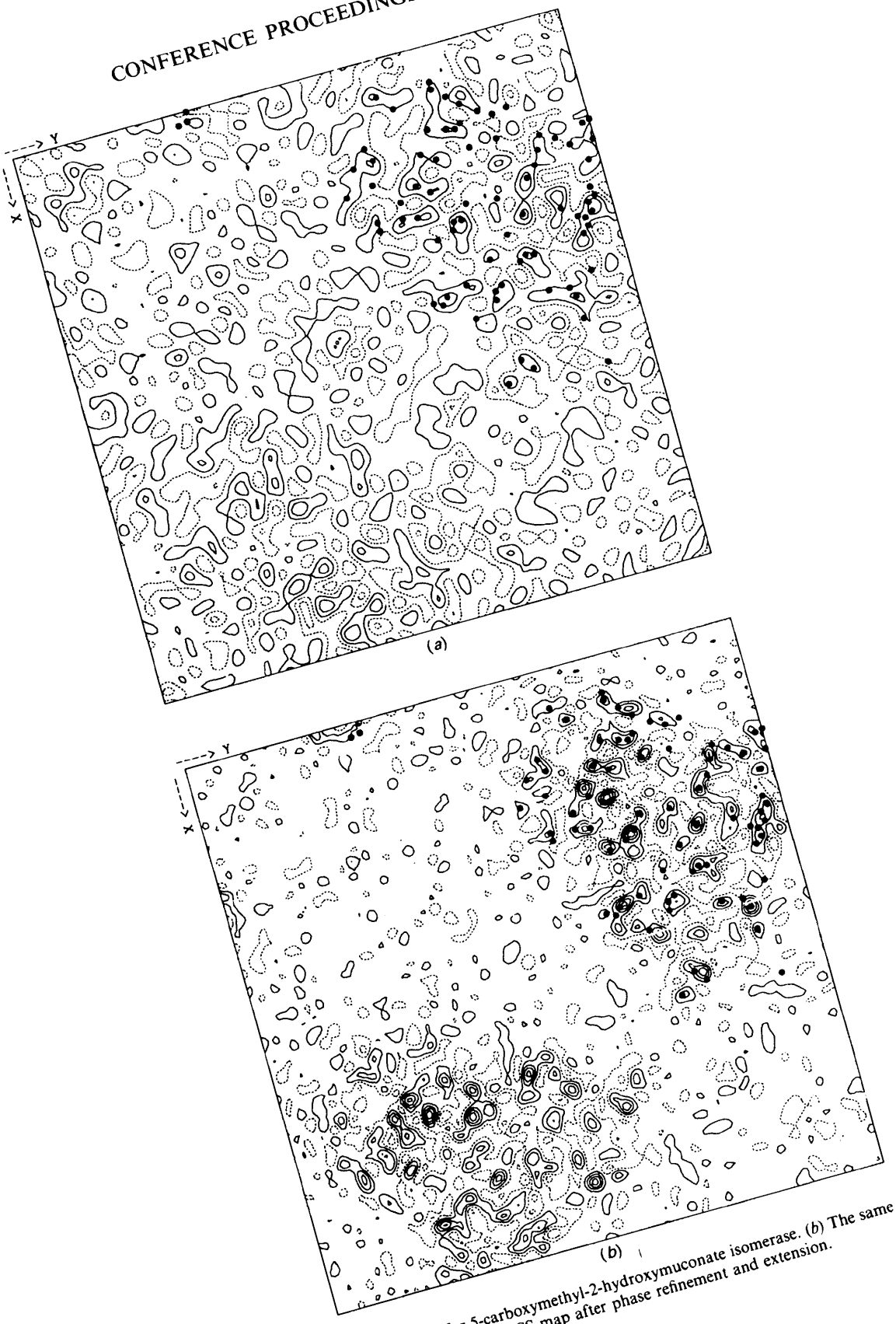
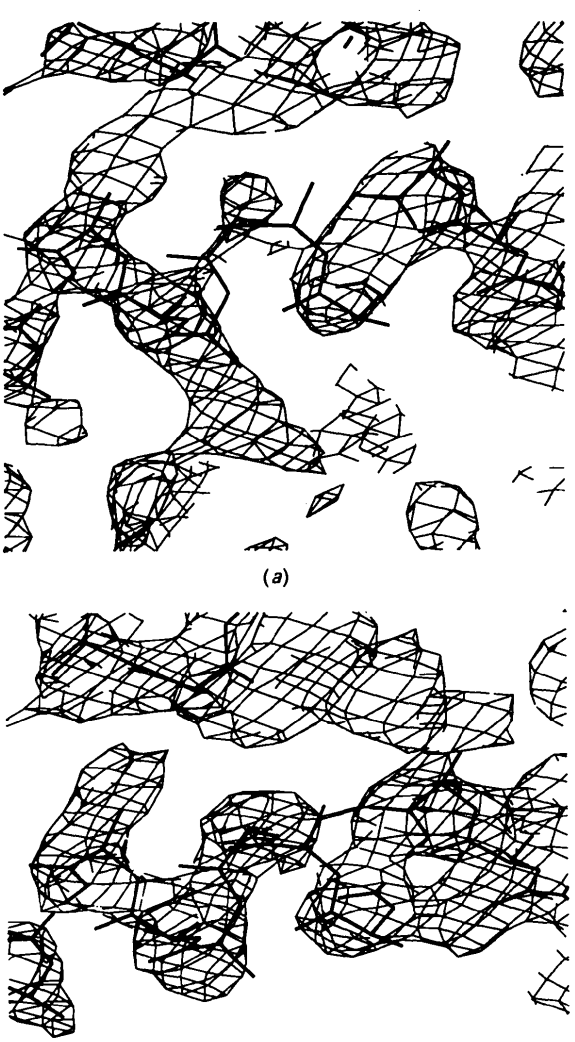


Fig. 3. (*a*) A section of the 3.7 Å MIR map for 5-carboxymethyl-2-hydroxymuconate isomerase. (*b*) The same section as in (*a*) but
2.1 Å HM/SF/NCS map after phase refinement and extension.

this map to extend the phases to the resolution limit of the native data. An improved map was generated from which it was possible to build a model of about 30% of the main chain, mainly sections of α-helix and β-sheet. The map is now being further improved with the aid of density modification and the partial structure information.

A three-dimensional density plot is shown in Fig. 4 for a typical region of the map, before and after improvement. The superimposed model is a section of α-helix, but no side chains have been constructed at this stage. In Fig. 4(*a*), it can be seen that the density corresponding to the α-helix is broken and that the remaining density is sufficiently unclear that fitting a helix would be highly speculative. In Fig. 4(*b*), the break in the density has been almost bridged and the remaining density is clearer. This is typical of the improvements which have transformed an uninterpretable map into one from which a structure determination can at least be attempted.



(*a*)

(*b*)

Fig. 4. (*a*) Part of the penicillin acylase MIR map. (*b*) The same part of the penicillin acylase HM/SF map.

6. Discussion

It has been shown that data from real and reciprocal space constraints on the electron density can be combined to form a practical algorithm for phase refinement. *SQUASH* represents a practical application of this method, incorporating Sayre's equation, solvent flattening, histogram matching, non-crystallographic symmetry and partial structure in a single procedure. The program has been applied without user intervention to a range of test structures and the results suggest that the simultaneous application of the various techniques is more effective than applying each of them separately.

Sayre's equation was very effective for phase refinement and extension in most of the test cases. However, if the initial map is poor, Sayre's equation becomes ineffective in phase refinement. Under these circumstances, is is better to apply density modification alone until a less noisy map is obtained. Sayre's equation also decreases in power as the solvent content increases, since it is only applicable to the molecular regions of the map. This suggests that a solvent-removal algorithm may improve its performance. It works best when high-resolution data (better than 3 Å) are available. It is also clear that the higher the resolution, the more accurate the phases of all reflections become.

Solvent flattening is poor for phase extension, but is good for phase refinement. Histogram matching is much better for phase extension and the combination of the two techniques can be very effective. They can be applied successfully to most structures.

Non-crystallographic symmetry averaging is useful for refining phases, but is very weak for phase extension without taking special precautions. The combination of non-crystallographic symmetry averaging with histogram matching and solvent flattening, however, is a very powerful technique for both phase refinement and extension.

Partial structure information as a density-modification technique is only a little better than its normal use in phase calculations.

The large improvement in electron-density maps produced by the combination of histogram matching, solvent flattening and non-crystallographic symmetry suggests that the strengths of the techniques are complementary. Each technique, when applied in isolation, will introduce systematic errors which are difficult to overcome when a different technique is

subsequently applied. This problem is greatly reduced when applying the techniques simultaneously and the combined process iterates much further towards the desired density map.

Work is continuing in the investigation of other techniques that can be incorporated into *SQUASH* and to try and improve the performance of Sayre's equation at low resolution. In particular, we wish to apply Sayre's equation only to the molecular region and to find an improved method for generating the θ function.

We wish to thank Professor G. G. Dodson and Mrs E. J. Dodson for the use of laboratory space and for help with the computing. We would also like to thank Drs D. Wigley and H. Duggleby for their help with some of the test structures. One of us (KDC) is grateful to the Rigaku Corporation of Japan for the provision of a research studentship.

References

ADAMS, M. J., BLUNDELL, T. L., DODSON, E. J., DODSON, G. G., VIJAYAN, N. M., BAKER, E. N., HARDING, M. M., HODGKIN, D. C., RIMMER, B. & SHEAT, S. (1969). *Nature (London)*, **224**, 491–495.

BAKER, E. N., BLUNDELL, T. N., CUTFIELD, J. F., CUTFIELD, S. M., DODSON, E. J., DODSON, G. G., HODGKIN, D. C., HUBBARD, R. E., ISAACS, N. W., REYNOLDS, C. D., SAKABE, N. & VIJAYAN, N. M. (1988). *Philos. Trans. R. Soc. London*, **319**, 369–456.

BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography*. London: Academic Press.

BRICOGNE, G. (1974). *Acta Cryst.* A**30**, 395–405.

DUGGLEBY, H., MOODY, P., HILL, C. P., DODSON, E. J., DODSON, G. G. & TROLLEY, S. P. (1992). In preparation.

FINZEL, B. C., CLANCY, L. L., HOLLAND, D. R., MUCHMORE, S. W., WATENPAUGH, K. D. & EINSPAHR, H. M. (1989). *J. Mol. Biol.* **209**, 779–791.

HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136–143.

MAIN, P. (1990). *Acta Cryst.* A**46**, 372–377.

OLDFIELD, T. J., DODSON, E. J., HUBBARD, R. E., JIANGSHENG, J., MURRAY-RUST, P., PAPIS, M. & TURKENBURG, M. (1992). In preparation.

READ, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

SAYRE, D. (1952). *Acta Cryst.* **5**, 60–65.

SAYRE, D. (1974). *Acta Cryst.* A**30**, 180–184.

ŠEVČÍK, J., BOTH, V., HALICKÝ, P. & ZELINKA, J. (1981). *Biologia (Bratislava)*, **36**, 235–239.

ŠEVČÍK, J. & ZELINKA, J. (1984). *Proceedings of the Fifth International Symposium on Metabolism and Enzymology of Nucleic Acids Including Gene Manipulation*, edited by J. ZELINKA & J. BALLAN, pp. 139–143. Bratislava: Veda.

SIM, G. A. (1959). *Acta Cryst.* **7**, 61–67.

SRINIVASAN, R. (1966). *Acta Cryst.* **20**, 143–144.

WANG, B. C. (1985). *Methods Enzymol.* **115**, 90–112.

WIGLEY, D., ROPER, D. I., DODSON, E. J., WILSON, K. S., DAUTER, Z. & DAVIES, G. D. (1992). In preparation.

WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

ZHANG, K. Y. J. (1989). PhD thesis, Univ. of York, England.

ZHANG, K. Y. J. & MAIN, P. (1990*a*). *Acta Cryst.* A**46**, 41–46.

ZHANG, K. Y. J. & MAIN, P. (1990*b*). *Acta Cryst.* A**46**, 377–381.