Acta Cryst. (1993). D**49**, 213–222

SQUASH – Combining Constraints for Macromolecular Phase Refinement and Extension

BY KAM Y. J. ZHANG

Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California at Los Angeles, Los Angeles, CA 90024-1570, *USA*

(*Received* 15 *June* 1992; *accepted* 21 *July* 1992)

Abstract

The constraints of correct electron-density distribution, solvent flatness, correct local shape of the electron density and equal molecules are combined in an integrated procedure for macromolecular phase refinement and extension. These constraints on electron densities are satisfied simultaneously by solving a system of non-linear equations. The electron-density solution is further filtered by a phase combination procedure. The non-crystallographic symmetry operations are refined by a rotation and translation space search and a least-squares minimization method, thereby reducing the chance of introducing systematic phase errors during averaging. The effect of each constraint on phase refinement and extension is examined. The constraints are found to work synergistically in phase improvement. Test results on 2Zn insulin are presented.

1. Introduction

The success of direct methods in solving the phase problem for small molecules demonstrates that constraints on electron density serve as a link between structure-factor amplitudes and phases. Exploitation of these constraints allows *ab initio* phase determination for small molecules from structure-factor amplitudes alone (Karle, 1986; Hauptman, 1986; Woolfson, 1987).

Direct methods for small molecules rely upon the availability of X-ray diffraction data at atomic resolution. The inequality relationship of Harker & Kasper (1948) introduced the constraint that electron-density values should be non-negative. A more effective method which maximizes $\int \rho^3 d\rho$ proposed by Cochran (1952) not only penalizes negative density but also constrains the positive density. Sayre's equation (Sayre, 1952) relates structure factors together by restricting the local shape of electron density and it has proven to be a very powerful method for *ab initio* phasing for small molecules.

The phase problem for macromolecules poses a different challenge. The marginal phase-probability relationship of a few structure-factor amplitudes and phases becomes less reliable due to the large number of atoms in the unit cell (Bricogne, 1984, 1988). Thus, the use of a joint-probability relationship or employing more marginal relationships, such as in the minimum principle (Hauptman, 1990), the Sayre tangent formula (Debaerdemaeker, Tate & Woolfson, 1985) and the maximum-entropy method (Bricogne, 1984, 1988; Prince, 1989), would be more effective for solving the macromolecular phase problem.

Crystals of macromolecules rarely diffract to atomic resolution. Thus, the positivity and atomicity constraints which are effective in solving the phase problem of small molecules are no longer strictly valid. However, one feature that crystals of most macromolecules have in common is the large solvent content. The solvent flattening technique based on this constraint has been very successful in macromolecular phase refinement and extension (Wang, 1985). When the same molecule occurs more than once in the asymmetric unit, this geometric redundancy can be exploited by non-crystallographic symmetry (NCS) averaging (Bricogne, 1974, 1976). The ideal density histogram also provides a constraint on the electron-density distribution which can be used to improve phases for macromolecules (Harrison, 1988; Lunin, 1988, Zhang & Main, 1990*a*). It was later demonstrated that density histograms can serve as a figure of merit to retrieve correct phase sets in *ab initio* phasing of macromolecules (Lunin, Urzhumtsev & Skovoroda, 1990).

The use of correct electron-density distribution, solvent flatness and correct local-density shape in a combined procedure – *SQUASH* – has been demonstrated to be a powerful method for macromolecular phase refinement and extension (Zhang & Main, 1990*b*). This work extends that method to include non-crystallographic symmetry averaging when more than one molecule occurs in the asymmetric unit. The constraint presented by this geometric redundancy is simultaneously satisfied together with other constraints in the solution to a system of non-linear equations. Test results on 2Zn insulin will be presented.

2. Constraints used in *SQUASH*

Characteristic features of the correct electron density can often be expressed as mathematical constraints on the electron density or the structure factors. Because the structure-factor amplitudes are known, these constraints restrict the values of the phases, and in favorable cases, this is sufficient to determine the phases directly.

The constraints used in this work are described below.

0907-4449/93/010213-10$06.00

© 1993 International Union of Crystallography

2.1. *Electron-density probability distribution – density histogram*

The density histogram of a map is the probability distribution of electron-density values. The comparison of the histogram for a given map with that expected for a good map can serve as a measure of quality. Furthermore, the initial map can be improved by adjusting density values in a systematic way to make its histogram match the ideal histogram.

The ideal histograms for protein structures depend on resolution and overall temperature factor, but they are independent of the particular structures (Zhang & Main, 1990*a*). The shape of the histogram is primarily based on the presence of atoms and their characteristic distances apart. This is true for all polypeptide structures. The correct histogram can be taken from a known structure (Zhang & Main, 1990*a*) or predicted by an analytical formula (Main, 1990*b*; Lunin & Skovoroda, 1991).

2.2. *Solvent flatness*

In macromolecular crystals, generally 30–80% of the volume is occupied by solvent. At medium or low resolution, the scattering of the solvent molecules is like that of a flat region of densities due to high thermal motion and disorder of the solvent.

The existence of a flat solvent region in a crystal places strong constraints on the structure-factor phases (Bricogne, 1974; Ward, Hendrickson & Klippenstein, 1975). Hendrickson & Lattmann (1970) showed that the smaller the molecular volume, the larger the number of structure factors that are algebraically related, *i.e.* the constraints are more powerful if a larger portion of the crystal volume is occupied by solvent.

2.3. *Local shape of the electron density*

Sayre (1952) pointed out that for equal and resolved atoms, the density distribution is equal to the squared density convoluted with an atomic shape function. This atomicity constraint can in principle be adapted to medium- or low-resolution data by defining groups of atoms that diffract as a single unit. Sayre (1972, 1974) also showed that the atomic shape function can be modified to compensate for data incompleteness. It has been successfully applied to the refinement and extension of the multiple isomorphous replacement (MIR) phases of rubredoxin from 2.5 to 1.5 Å resolution.

2.4. *Equal molecules*

When the same molecule occurs more than once in the asymmetric unit, the distinct molecules usually have a similar three-dimensional structure. These molecules can be related by non-crystallographic point-group symmetry or general rotation and translation operations. It can be assumed that the corresponding density values are approximately equal between the related molecules. This places a strong constraint on the density values at equivalent positions among the molecules. Phase relationships arising from this constraint can be formulated either in reciprocal space (Rossmann & Blow, 1962; Main & Rossmann, 1966; Main, 1967) or in real space (Argos, Ford & Rossmann, 1974; Bricogne, 1974; Buehner, Ford, Moras, Olsen & Rossmann, 1974). The equivalence and relative effectiveness of the real and reciprocal space approaches were discussed by Bricogne (1974). A successful phase refinement and extension, or even *ab initio* phasing in favorable cases, can be achieved by exploiting the geometric redundancies presented by equal molecules (Tsao, Chapman & Rossmann, 1992; Chapman, Tsao & Rossmann, 1992).

3. Methods

The phasing power of a method increases with the number of independent constraints employed, the number of density points affected and the magnitude of changes imposed on the electron density. It also depends on the physical nature and accuracy of the constraint and how rigorously the constraint is applied.

The constraints described above are implemented in the following methods for macromolecular phase refinement and extension.

3.1. *Histogram matching*

The matching of histogram $P'(\rho)$ to $P(\rho)$ is achieved by finding the upper bound ρ_i' corresponding to ρ_i in the following equation (Zhang & Main, 1990*a*),

$$\int_{\rho_0}^{\rho_i} P(\rho)\mathrm{d}\rho = \int_{\rho_0'}^{\rho_i'} P'(\rho)\mathrm{d}\rho \qquad (1)$$

where ρ_0 and ρ_0' are the minimum densities in the histograms $P(\rho)$ and $P'(\rho)$ respectively.

If the sampling i is sufficiently fine, a linear transformation between ρ_i and ρ_i' can be used,

$$\rho_i' = a_i\rho_i + b_i \qquad (2)$$

where a_i and b_i are the scale and shift respectively.

Note that histogram matching applies a minimum and a maximum value to the electron density, imposes the correct mean and variance, and defines the entropy of the new map. The order of electron-density values also remains unchanged after histogram matching.

The density histogram depends upon the overall temperature factor as mentioned previously. However, it was found that the best phases are obtained after removing the effect of temperature from the F's (Zhang & Main, 1990*a*). Thus, if amplitudes sharpened in this way are always used, it becomes unnecessary to change the histogram according to the temperature factor of different structures.

3.2. *Solvent flattening*

The constraint of solvent flatness is implemented by identifying the molecular boundaries and replacing the densities in the solvent region by their mean density value. The molecular boundary is located by an automated procedure proposed by Wang (1985) and modified by Leslie (1987).

3.3. *Sayre's equation*

For equal and resolved atoms, Sayre's equation (Sayre, 1952) relates the structure factors exactly as

$$F(\mathbf{h}) = [\theta(\mathbf{h})/V] \sum_{\mathbf{k}} F(\mathbf{k})F(\mathbf{h}-\mathbf{k}) \qquad (3)$$

where $\theta(\mathbf{h}) = f(\mathbf{h})/g(\mathbf{h})$ is the ratio of scattering factors of real and squared atoms and V is the unit-cell volume.

Because electron-density constraints are more easily expressed in real space than in reciprocal space, it is convenient to express Sayre's equation in terms of the electron density,

$$\rho(\mathbf{n}) = (V/N) \sum_{\mathbf{m}} \rho^2(\mathbf{m})\psi(\mathbf{n}-\mathbf{m}) \qquad (4)$$

where

$$\psi(\mathbf{n}-\mathbf{m}) = (1/V) \sum_{\mathbf{h}} \theta(\mathbf{h})\exp[-2\pi i\mathbf{h}(\mathbf{n}-\mathbf{m})] \qquad (5)$$

which states that the convolution of squared electron density with a shape function produces the original electron density. It can be seen from (4) that Sayre's equation puts constraints on the local shape of electron density. The shape function is represented by $\psi(\mathbf{n})$ in (5), which is the Fourier transformation of $\theta(\mathbf{h})$.

The scale factor $\theta(\mathbf{h})$ is sensitive to resolution. Its shape can be predicted from the atomic scattering factors at atomic resolution. A more satisfactory method for non-atomic resolution work is to determine $\theta(\mathbf{h})$ as a function of $\sin\theta/\lambda$ by spherically averaging the ratio between $F(\mathbf{h})$ and the structure factors $G(\mathbf{h})$ from the squared electron density. A least-squares curve is fitted on the experimental data which takes the form

$$\theta(s) = K\exp(-Bs^2). \qquad (6)$$

where $s = \sin\theta/\lambda$. The $\theta(\mathbf{h})$ is taken from the spherically averaged $\theta(s)$.

Alternatively $\theta(\mathbf{h})$ can be taken from the spherically averaged ratio, $F(\mathbf{h})/G(\mathbf{h})$, from the electron-density map at the given resolution and extrapolated to higher resolution by (6) if phase extension is required (Cowtan & Main, 1993).

3.4. *Non-crystallographic symmetry refinement*

The initial NCS operation obtained from rotation and translation functions (Rossmann & Blow, 1962; Crowther & Blow, 1967) or heavy-atom positions can be fine-tuned by a density space R-factor search in the six-dimensional rotation and translation space. The density space R factor is

$$R = \sum_{\mathbf{r}} |\rho(\mathbf{r}) - \rho(\mathbf{r}')| / \sum_{\mathbf{r}} |\rho(\mathbf{r}) + \rho(\mathbf{r}')| \qquad (7)$$

as defined by Brändén & Jones (1990), where $\mathbf{r}' = \mathbf{\Omega r}$ and $\mathbf{\Omega}$ represents the NCS operator.

The search rate is increased by using only a representative subset of grid points. Grid points which have density values above a specified threshold and are greater than a specified distance from each other are selected within the subunit where the NCS operation holds. The NCS operation is systematically altered to find the lowest density space R factor for the selected subset of grid points.

The translation search rate can be increased by using three fast Fourier transforms according to the convolution theorem. Then, the six-dimensional rotation and translation space search can be reduced to a three-dimensional rotation space search with three fast Fourier transforms.

The NCS operation solution from the six-dimensional rotation and translation search can be further refined by a least-squares procedure.

If $\rho(\mathbf{r})$ is related to $\rho(\mathbf{r}')$ by the NCS operation $\mathbf{\Omega}$,

$$\rho(\mathbf{r}') = \rho(\mathbf{\Omega r}) \qquad (8)$$

where $\mathbf{r}' = \mathbf{\Omega r}$. Here, $\mathbf{\Omega}$ is a function of ω, $\mathbf{\Omega} = f(\omega)$, where $\omega = (\alpha, \beta, \gamma, t_x, t_y, t_z)$ representing the rotation and translation components of the NCS operation.

The residual,

$$\varepsilon(\mathbf{r}) = \rho(\mathbf{r}) - \rho(\mathbf{\Omega r}) \qquad (9)$$

can be minimized by a least-squares formula of the form,

$$(\partial\rho/\partial\omega)^T(\partial\rho/\partial\omega)\Delta\omega = (\partial\rho/\partial\omega)^T\varepsilon(\mathbf{r}) \qquad (10)$$

where $\Delta\omega$ is the shift to the NCS operation. Here, $\partial\rho/\partial\omega = (\partial\rho/\partial\mathbf{r})(\partial\mathbf{r}/\partial\omega)$ where the partial derivative $\partial\rho/\partial\mathbf{r}$ is calculated by fast Fourier transforms and $\partial\mathbf{r}/\partial\omega$ is evaluated analytically.

3.5. *Averaging*

Averaging is carried out as proposed by Bricogne (1976) except that the double-sorting procedure is avoided by storing the whole map in the memory. Every grid point in the unit cell is mapped to the grid point within the subunit mask where the NCS holds and the NCS operation is subsequently applied. The mapping procedure gives the appropriate crystallographic or non-crystallographic operation which could transform each grid point to the subunit

mask. Thus, the mask for partitioning the protein from solvent can be derived from this mapping.

The average density value for each NCS-related grid point is determined. The value assigned to each grid point can be adjusted toward the mean density according to a weight. If $\bar{\rho}$ represents the mean density from the average of ρ_i with $i = \{1, n\}$, then the modified density is taken as

$$\rho_i' = \rho_i + w(\bar{\rho} - \rho_i) \tag{11}$$

where w ranges from 0 to 1. The weight can be used as a function of map resolution or spatial distribution of grid points.

3.6. *Phase combination*

Having obtained the modified electron-density map, the structure-factor amplitudes and phases can be calculated from its inverse Fourier transform. It is desirable to weight each structure factor according to the accuracy of its phase. Not only will the noise in the final electron-density map be reduced, but the use of weighting during the iteration will also make the process less noise sensitive, hence widening its radius of convergence. It will also allow the well phased reflections to correct the poorly phased ones, while minimizing the detrimental effect of the initially poorly determined phases.

The MIR phase-probability distribution is given by Blow & Crick (1959). The probability distribution for phases calculated from the modified map is determined using Sim's weighting scheme (Sim, 1959) as adapted by Bricogne (1976). The phases are combined by multiplying their respective phase probabilities (Bricogne, 1976). This multiplication of phase probabilities is simplified by adding the coefficients that code for phase probabilities (Hendrickson & Lattman, 1970).

3.7. *Combining constraints and solution to the system of constraint equations*

Density modifications, such as solvent flattening, histogram matching and NCS averaging, put constraints on the density map, which can be expressed as,

$$\rho(\mathbf{n}) = H(\mathbf{n}) \tag{12}$$

where $H(\mathbf{n})$ represents the modified map.

For the electron density to satisfy (12) and (4) simultaneously, we need to solve the following system of equations,

$$\begin{cases} (V/N)\sum_{\mathbf{m}} \rho^2(\mathbf{m})\psi(\mathbf{n}-\mathbf{m}) - \rho(\mathbf{n}) = 0 \\ wH(\mathbf{n}) - w\rho(\mathbf{n}) = 0. \end{cases} \tag{13}$$

with $\mathbf{n} = \{1, N\}$, where w is the relative weight between (4) and the constraints from density modification (12).

Equations (13) represent a system of non-linear simultaneous equations with as many unknowns, $\rho(\mathbf{n})$, as grid points, N, in the asymmetric unit of the map and twice as many equations as unknowns. The functions $H(\mathbf{n})$ and $\psi(\mathbf{n})$ are both known. The least-squares solution, using either the full-matrix or the diagonal approximation, is obtained using the Newton-Raphson technique as described by Main (1990*a*).

4. Procedure

The general procedure of phase refinement and extension is illustrated in Fig. 1 and described by the following steps:

(1) The mask of the subunit which corresponds to one of the molecules related by NCS in the asymmetric unit should be identified by visual inspection of the initial MIR map or other methods. This mask will be referred to as the subunit mask later on.

(2) The initial NCS operation such as that obtained from rotation and translation functions or heavy-atom positions is fine-tuned by the density space *R*-factor search method. It can be subsequently refined by the least-squares minimization method.

(3) The electron densities within the protein region are averaged according to the NCS operations refined above. Different weights can be given with respect to the resolution of the map and how closely the subunits follow the NCS.

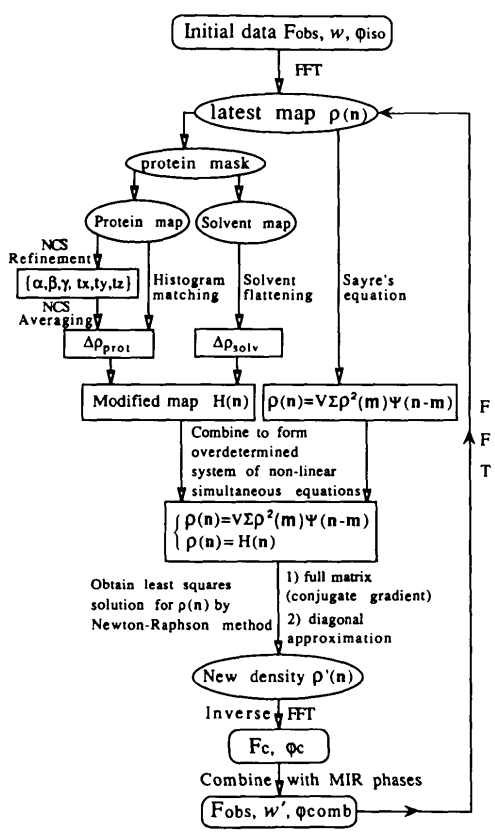

Fig. 1. Diagram of the *SQUASH*ing procedure.

(4) The histogram is computed from the averaged electron densities in the protein region and the map is subsequently transformed by matching its histogram to the ideal histogram. The reason that the averaging is performed before the histogram matching is that the grid points which have the same density value will still have the same density value after histogram matching. Thus, the averaged density map after histogram matching satisfies the NCS constraint and the histogram constraint simultaneously; whereas the reverse process would not guarantee that.

(5) The solvent region is set to its mean density. The modified map from steps 3, 4 and 5 gives the density modification equation (12).

(6) Sayre's equation (4) is set up. The shape function $\theta(\mathbf{h})$ is computed by spherical averaging at the resolution of phase refinement and extrapolated by an exponential model to the extended resolution.

(7) The system of non-linear simultaneous equations is solved by the conjugate-gradient method using fast Fourier transforms as proposed by Main (1990*a*). Different weights can be given to (13) to reflect the different phasing power of Sayre's equation and other constraints at a given resolution of a particular system.

(8) New structure-factor amplitudes and phases are calculated from the inverse Fourier transform of the modified map. The Sim weight is calculated for each new structure factor. The new phases and Sim weights are combined with the MIR phases to produce the combined phases and figure of merit.

(9) For phase extension, a new set of observed structure factors are selected and added to the starting set. Their initial phases are taken from the inverse Fourier transform of the modified map. New reflections can be added either in increasing resolution shells or in decreasing order of structure-factor amplitudes with equal increments of the $\sum F^2$ for the added reflections so as to give the same amount of perturbation to the map. A modified procedure which includes new reflections in order of resolution-weighted amplitudes strikes a compromise between the two choices above (Cowtan & Main, 1993). Here it appears that the best procedure is to add new reflections in order of the resolution weighted structure-factor amplitudes, but with decreasing $\sum F^2$ in each step.

(10) A new map is calculated from the combined structure-factor amplitudes and phases and the whole procedure is reiterated until convergence.

5. Test results

The known structure of 2Zn pig insulin (Baker *et al.*, 1988) was chosen as a test for the method. The space group is *R*3 with two insulin molecules of 51 amino-acid residues in the asymmetric unit. There is 30% solvent in the crystal. The amplitudes used in the calculations were the observed *F*'s which had the overall temperature factor removed.

The NCS which relates the dimer of insulin in the asymmetric unit was identified by the program *ALMN* (SERC Daresbury Laboratory, 1984) using the Crowther fast rotation function (Crowther, 1972) which gave a self rotation solution at Eulerian angles (0, 180, 330.4°). A mask of the insulin monomer was created by building dummy atoms into the MIR map, followed by averaging the electron-density map calculated from the dummy atoms using a 10 Å sphere. The initial rotation angles (0, 180, 330.4°) were refined by the density space *R*-factor search and least-squares minimization method to final values of (0, 179.6, 331.5°). The translational part of the NCS operation was found to be (0.007, -0.166, -0.432 Å) in the orthogonal coordinate system. The density space *R* factor was reduced from 34.8 to 27.0% after NCS refinement.

MIR phases to 3.0 Å resolution were used to calculate the starting map. The electron-density map was averaged according to the refined NCS operation. The averaged map was modified by histogram matching and the solvent region was set to its mean density. A new electron-density map which satisfied both Sayre's equation and the density-modification constraints was derived from a least-squares solution. This constitutes one cycle of phase improvement. The phases were extended from 3.0 to 2.0 Å in ten steps. The new reflections were added in order of weighted structure-factor amplitudes. Subsequently, the phases were extended from 2.0 to 1.5 Å, the resolution limit of the native insulin data set, in ten steps.

In order to address the question of how each constraint contributed to the map improvement, the phase refinement and extension from 3.0 to 1.5 Å were carried out using each individual constraint or selective combinations of them. The results are compared with the MIR phases as shown in Figs. 2 and 3. A summary of these results is listed in Table 1.

It can be seen from Fig. 2 and Table 1 that solvent flattening refined the phases well but phase extension was not effective. This is probably due to the low solvent content. Histogram matching did well in phase refinement, but it also extended the phases well to about 2.0 Å. The effectiveness in phase extension is due to the fact that the correct histogram for the extended resolution explicitly modifies the map to a higher resolution. Averaging performed well for refinement but upon phase extension it produced no better than random phases, probably because there is only twofold redundancy so the phasing power is relatively low. Moreover, some side-chain atoms do not follow the NCS operation and this becomes more significant at higher resolution. Sayre's equation did not refine the phases but scored the best in phase extension. Since the constraint in Sayre's equation is less valid at lower resolution and becomes more valid with increasing resolution, moreover, Sayre's equation explicitly relates structure factors of higher resolution with that of lower resolution, and is therefore most effective for phase extension.

It can be seen from Fig. 3 and Table 1 that when histogram matching was combined with solvent flattening, there was further improvement in phase refinement and the phase extension was greatly improved over the results of

Table 1. *Comparison of phase refinement and extension using various constraints*

Constraints	Mean phase error $\langle\Delta\varphi\rangle$ (°) Refinement (∞-3.0 Å)	Extension (3.0-1.5 Å)
MIR	46.2	—
SF	42.8	89.6
HM	42.9	84.6
NCS	43.8	89.6
SAYR	45.9	75.5
SF+HM	41.0	78.9
SF+HM+NCS	38.6	76.3
SF+HM+SAYR	39.5	65.5
SF+HM+NCS+SAYR	38.1	61.7

Abbreviations: MIR, multiple isomorphous replacement; SF, solvent flattening; HM, histogram matching; NCS, non-crystallographic symmetry averaging; SAYR, Sayre's equation.

using solvent flattening or histogram matching alone. The addition of Sayre's equation combined the power of phase extension from Sayre's equation and the phase-refinement power of histogram matching and solvent flattening. When NCS averaging was used together with histogram matching and solvent flattening, the phases were refined further and the phase extension was improved over the lower resolution range up to 2.0 Å, but from 2.0 to 1.5 Å it gave the same result as using histogram matching and solvent flattening. The combination of all the constraints including histogram matching, solvent flattening, Sayre's equation and averaging gave the best results both in phase refinement and extension. The initial MIR phases of 1677 reflections at 3.0 Å were refined from a mean phase error of 46.0-38.1°. The extended phases for 10 729 reflections from 3.0 to 1.5 Å have a mean phase error of 61.7°.

The unweighted mean phase error does not take into account the different contribution to the electron density from structure factors of different amplitudes. Since we are interested in the improvement of the electron-density maps, a more pertinent measure would be the correlation coefficient between the modified and the correct map. Moreover, it is the protein part of the density we are concerned about and a residue-by-residue correlation coefficient would be more indicative of the regions where improvement were made. A map which encodes the electron-density value, the residue number and whether it belongs to the main chain or side chain for each grid point is calculated from the atomic model. The residue-by-residue correlation coefficient can be calculated from this encoded map and the map to be evaluated, in a single step (Dod-

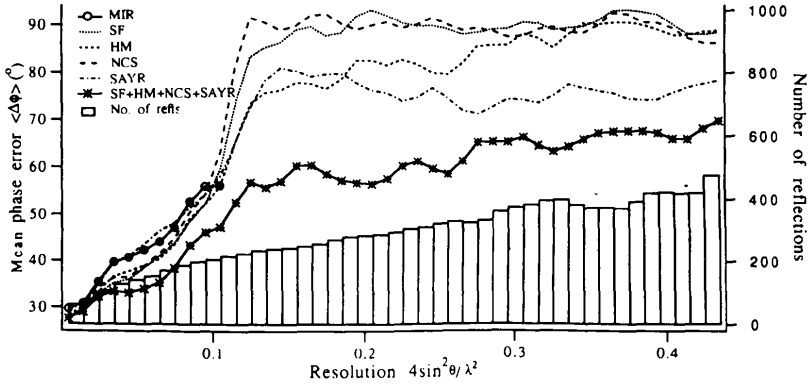

Fig. 2. Mean phase error as function of resolution using individual constraints for phase refinement and extension as compared with using all the constraints simultaneously. The mean phase error is the unweighted average phase difference between the phases in question and the correct phases calculated from the atomic model, *i.e.* $\Delta\varphi = \langle\varphi - \varphi_c\rangle$. The ordinate is the resolution. The left-hand abscissa is the mean phase error to which those curves correspond. The right-hand abscissa is the number of reflections and corresponds to the histogram bars.

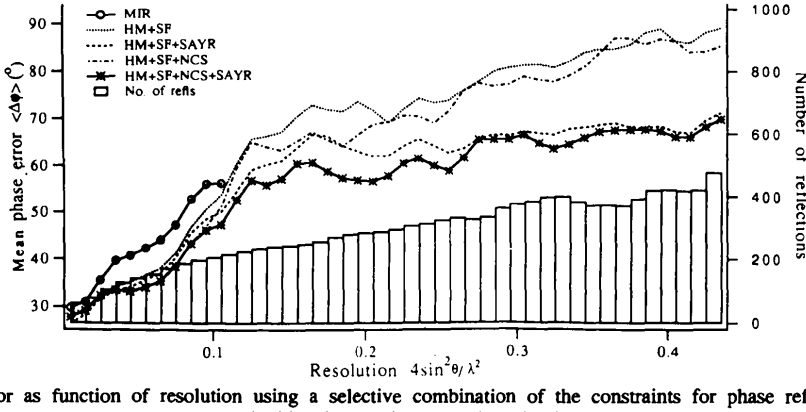

Fig. 3. Mean phase error as function of resolution using a selective combination of the constraints for phase refinement and extension as compared with using all the constraints simultaneously.

son, 1988). The total volumes occupied by the main-chain and side-chain atoms are 23 and 47% respectively given a radius of 2.5 Å for each atom. Figs. 4 and 5 show the correlation coefficients for the main-chain and side-chain atoms respectively. Those residues with side-chain correlation coefficients equal to zero are glycines. The majority of residues had significant improvements in the correlation coefficient for both main-chain and side-chain atoms. The correlation coefficient for main-chain atoms was improved from 0.62 to 0.78 overall. The correlation coefficient for side-chain atoms was increased from 0.56 to 0.70 on average. This resulted in a much improved map.

6. Discussion

The refinement of NCS operations prior to averaging minimized the chance of introducing systematic errors into the phases during the averaging process due to inaccuracies in the NCS operators. The algorithms used in the refinement of NCS operations are general and can be used to refine any NCS operation. The density space *R*-factor search method can be used to locate the NCS *ab initio* from the electron-density map. The disadvantage as compared with the rotation and translation function is that initial phases are required. However, it should be more sensitive since there is no vector overlap and less degeneracy than in the Patterson space. When the NCS axis is approximately parallel to a crystallographic axis of the same degree of rotation, the solution of the rotation function is hidden under the solution for the crystallographically related vector set. However, the density space *R*-factor search can reveal such a NCS operation. It can also be used when no model structure is known, in which case only the rotation part of the NCS can be derived from the rotation function, the translational components can be found by a density space *R*-factor search. The least-squares refinement of NCS operations after each cycle of density modification ensures that the differences between the NCS-related subunits are primarily due to errors in the MIR phases rather than inaccuracies in NCS operations.

The averaging procedure adopted can deal with subunits related by a general rotation and translation operation provided that the mask of a single subunit is known. The disadvantage is that the mask cannot be updated automatically after each cycle of refinement. Another way of implementing the procedure would be to use only

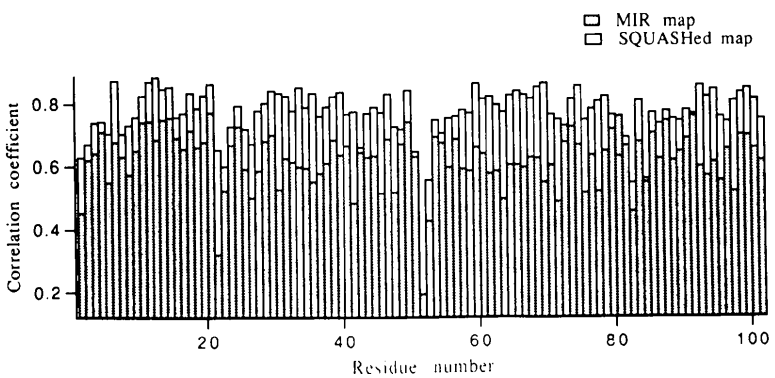

Fig. 4. A residue-by-residue correlation coefficient for the main-chain atoms. The correlation coefficient is the Pearson linear correlation of the expression: $C = (N\sum \rho_i\rho_i' - \sum \rho_i \sum \rho_i')/\{[N\sum \rho_i^2 - (\sum \rho_i)^2][N\sum \rho_i'^2 - (\sum \rho_i')^2]\}^{1/2}$ where ρ_i and ρ_i' are the electron-density values of the two maps at the ith grid point and N is the total number of grid points in the map.

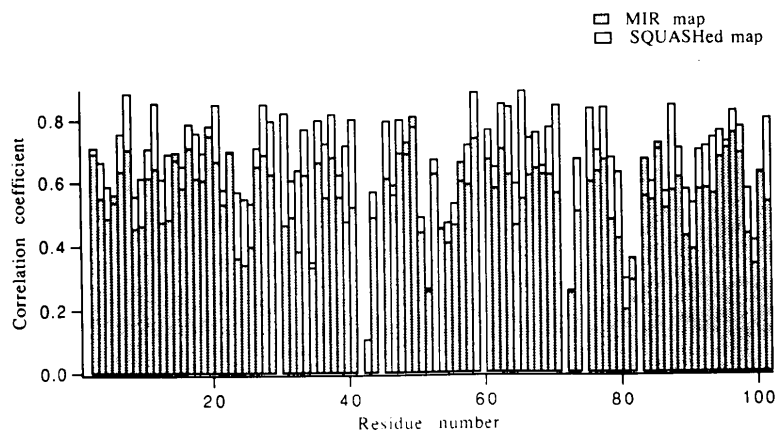

Fig. 5. A residue-by-residue correlation coefficient for the side-chain atoms.

the subunit mask for averaging; the solvent and protein partition mask can still be calculated automatically by the procedure of Wang (1985). In this way, the mask can be updated after every cycle of map improvement. Moreover, parts of the protein which do not follow the NCS can be excluded from averaging.

From the comparison of the phase improvement using different constraints, we can get a realistic estimate of how each constraint behaves for the phase improvement of a particular structure and how it changes as a function of resolution. The relative effectiveness of constraints can be used as a weight to balance the role of each constraint when used simultaneously. Since Sayre's equation, for instance, is effective for phase extension but not for phase refinement at lower resolution, a relatively low weight can be given to Sayre's equation at the start of phase refinement and the weight can be gradually increased with increasing resolution. It can be seen that the NCS averaging did well in phase refinement but it was not effective in phase extension at high resolution. One way of using this information is to take the full suggested modification to the electron density as indicated by the NCS averaging at low resolution, but only a partial suggested modification at high resolution.

The fact that the best results were obtained when all the constraints were combined together indicates that each constraint contains some degree of independent phasing information. Moreover, it also suggests that the constraints are complementary. This can be seen from the effect of the histogram matching when combined with Sayre's equation. Sayre's equation depends critically on the scaling of the structure-factor amplitudes. The scale and overall temperature factor estimated from Wilson statistics (Wilson, 1942) are usually not accurate enough to put the observed structure-factor amplitudes on absolute scale, especially when only medium- or low-resolution data are available. Sayre's equation, when used alone, tends to diverge since the scaling errors in amplitudes accumulate during iteration. The addition of histogram matching helps to constrain the structure-factor amplitudes by applying a minimum, maximum, mean and variance to the modified electron density. This stabilizes the system and makes Sayre's equation more effective.

The distribution of mean phase error as a function of structure-factor amplitudes revealed that stronger reflections had larger phase improvements. Figs. 6 and 7 show the mean phase errors as a function of structure-factor amplitudes and normalized structure-factor amplitudes, *E*'s.

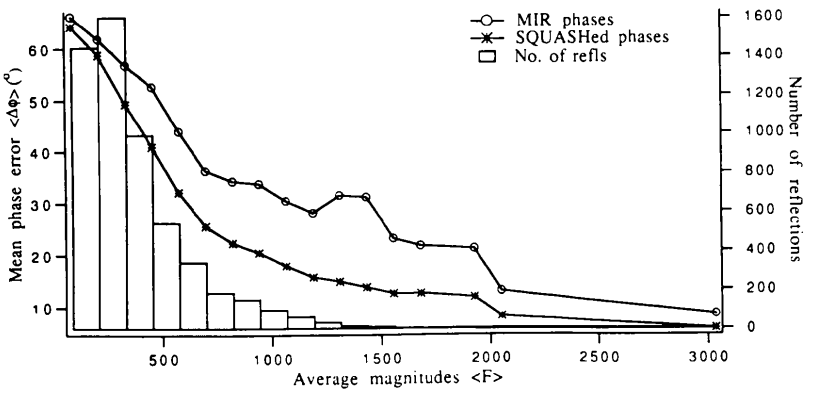

Fig. 6. Mean phase error as a function of structure-factor amplitudes at 2.0 Å.

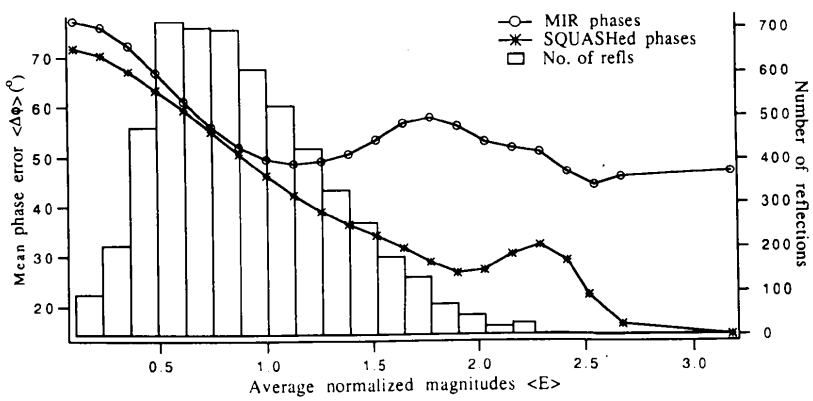

Fig. 7. Mean phase error as a function of normalized structure-factor amplitudes at 2.0 Å.

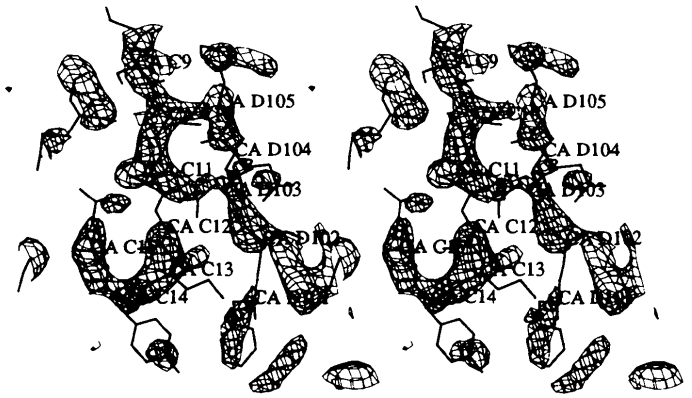

Fig. 8. MIR stereoscopic map at 3.0 Å resolution. The map is contoured at the 1.0σ level.

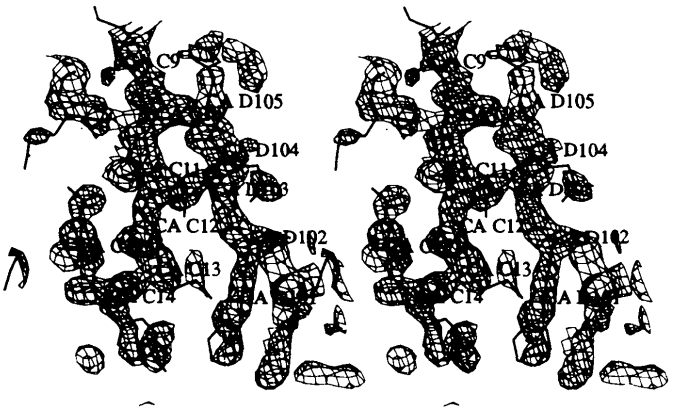

Fig. 9. Stereoscopic *SQUASH*ed map at 1.5 Å resolution. The map is contoured at the 1.0σ level.

For example, the phases for the top 1000 *F*'s have a mean phase error of 24.9° as compared with the mean phase error of 37.0° for the MIR phases. The mean phase error for the 527 reflections with *E*'s larger than 1.5 was improved from 55.0 to 30.8°. Interestingly, the phases for reflections with small *E*'s were improved more than those with medium *E*'s.

A portion of the electron-density map shown in Figs. 8 and 9 exemplifies the map improvement. Fig. 8 shows the MIR map at 3.0 Å resolution for residues from 9 to 15 in chain *C* and residues 101 to 105 in chain *D*. It can be seen that the main chain is broken between Cys 11 and Ser 12, Phe 101 and Val 102, Asn 103 and Gln 104. There is little density for the side chains of Ser 12, Leu 13, Tyr 14, Gln 15, Phe 101 and Gln 104. The same portion of the map after phase refinement at 3.0 Å resolution and subsequent extension to 1.5 Å resolution using all the constraints in *SQUASH* is shown in Fig. 9. It is evident that there is a considerable improvement in the overall quality of the map. The main chains between Cys 11 and Ser 12, Phe 101 and Val 102, Asn 103 and Gln 104 are connected. New densities appeared for the side chains of Ser 12, Leu 13, Tyr 14, Gln 15, Phe 101 and Gln 104. Moreover, the densities for the main-chain carbonyl atoms are clearly visible. This shows the high resolution and quality of the *SQUASH*ed map. The main-chain connectivity and side-chain densities are very important for the map interpretation and it is clear that this is significantly improved by the *SQUASH* process.

I would like to thank Professor G. G. Dodson for permission to use the 2Zn insulin data and Dr P. Main for many valuable discussions. I am very grateful to Professor D. Eisenberg for providing laboratory facilities and encouragement. This work was supported by NIH grant GM31299.

References

ARGOS, P., FORD, G. C. & ROSSMANN, M. G. (1974). *Acta Cryst.* A**31**, 499-506.

BAKER, E. N., BLUNDELL, T. L., 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, M. (1988). *Philos. Trans. R. Soc. London Ser. B*, **319**, 369-456.

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802

BRÄNDÉN, C. I. & JONES, A. (1990). *Nature (London)*, **343**, 687-689.

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

BRICOGNE, G. (1976). *Acta Cryst.* A**32**, 832-847

BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410-445.
BRICOGNE, G. (1988). *Acta Cryst.* A**44**, 517-545.
BUEHNER, M., FORD, G. C., MORAS, D., OLSEN, K. W. & ROSSMANN, M. G. (1974). *J. Mol. Biol.* **90**, 25-49.
CHAPMAN, M. S., TSAO, J. & ROSSMANN, M. G. (1992). *Acta Cryst.* A**48**, 301-312.
COCHRAN, W. (1952). *Acta Cryst.* **5**, 65-67.
COWTAN, K. D. & MAIN, P. (1993). *Acta Cryst.* D**49**, 148-157.
CROWTHER, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. ROSSMANN, pp. 173-178. New York: Gordon & Breach.
CROWTHER, R. A. & BLOW, D. M. (1967). *Acta Cryst.* **23**, 544-548.
DEBAERDEMAEKER, T., TATE, C. & WOOLFSON, M. M. (1985). *Acta Cryst.* A**41**, 286-290.
DODSON, E. (1988). *Proceedings of the Study Weekend Held at Daresbury Laboratory*, pp. 73-87. Warrington: SERC Daresbury Laboratory.
HARKER, D. & KASPER, J. S. (1948). *Acta Cryst.* **1**, 70-75.
HARRISON, R. W. (1988). *J. Appl. Cryst.* **21**, 949-952.
HAUPTMAN, H. (1986). *Science*, **233**, 178-183.
HAUPTMAN, H. (1990). In *Proceedings of the 16th International School of Crystallography*, Erice, Italy.
HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136-143.
KARLE, J. (1986). *Science*, **232**, 837-843.
LESLIE, A. G. W. (1987). *Acta Cryst.* A**43**, 134-136.
LUNIN, V. YU. (1988). *Acta Cryst.* A**44**, 144-150.
LUNIN, V. YU. & SKOVORODA, T. P. (1991). *Acta Cryst.* A**47**, 45-52.
LUNIN, V. YU., URZHUMTSEV, A. G. & SKOVORODA, T. P. (1990). *Acta Cryst.* A**46**, 540-544.
MAIN, P. (1967). *Acta Cryst.* **23**, 50.
MAIN, P. (1990*a*). *Acta Cryst.* A**46**, 372-377.
MAIN, P. (1990*b*). *Acta Cryst.* A**46**, 507-509.
MAIN, P. & ROSSMANN, M. G. (1966). *Acta Cryst.* **21**, 67-72.
PRINCE, E. (1989). *Acta Cryst.* A**45**, 200-203.
ROSSMANN, M. G. & BLOW, D. M. (1962). *Acta Cryst.* **15**, 24-31.
SAYRE, D. (1952). *Acta Cryst.* **5**, 60-65
SAYRE, D. (1972). *Acta Cryst.* A**28**, 210-212.
SAYRE, D. (1974). *Acta Cryst.* A**30**, 180-184.
SERC Daresbury Laboratory (1986). *CCP4. A Suite of Programs for Protein Crystallography.* SERC Daresbury Laboratory, Warrington, England.
SIM, G. A. (1959). *Acta Cryst.* **12**, 813-815.
TSAO, J., CHAPMAN, M. S. & ROSSMANN, M. G. (1992). *Acta Cryst.* A**48** 293-301.
WANG, B. C. (1985). *Methods Enzymol.* **115**, 90-112.
WARD, K. B., HENDRICKSON, W. A. & KLIPPENSTEIN, G. L. (1975). *Nature (London)*, **257**, 818.
WILSON, A. J. C. (1942). *Nature (London)*, **150**, 151.
WOOLFSON, M. M. (1987). *Acta Cryst.* A**43**, 593-612.
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.