Acta Cryst. (1995). D**51**, 990–1002

Phase Improvement by Cross-Validated Density Modification

BY ALISON L. U. ROBERTS AND AXEL T. BRÜNGER

The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, *USA*

(*Received* 19 *December* 1994; *accepted* 15 *March* 1995)

Abstract

Solvent flattening is a useful constraint for the early stages of crystallographic structure determination. However, sometimes it fails to produce significant improvement of poor experimental or molecular-replacement phases. This often occurs as a result of incorrect parameterization. In addition, the potential of overfitting or misinterpretation of the data exists. We have implemented a cross-validated (or free) R value in order to reduce this risk. The free R value was calculated between the experimental $\mathbf{F}_{obs}(\mathbf{h})$ and the calculated structure factors, $\mathbf{F}_{sf}(\mathbf{h})$, obtained by inverse Fourier transformation of the solvent-flattened electron density. Because of the sensitivity of the free R value to the test set selection at low resolution complete cross-validation may be required. The reliability of this approach was assessed by examining the correlation between the free R value and the known phase errors for two test cases. A high correlation was found upon variation of the extent of negative density elimination, figure of merit estimation, and the relative weighting in the phase combination procedure. The free R value can be used to optimize parameters of density-modification procedures when independent phase error estimates are unavailable.

1. Introduction

Phase information is required in addition to the measured diffraction intensities to solve a crystal structure by X-ray crystallography. For macromolecular structures, *ab initio* phasing by direct methods has not yet succeeded due to the limited amount of information contained in the structure-factor amplitudes (Giacovazzo, Siliqi & Ralph, 1994). Instead, initial phases are usually obtained from experimental methods or molecular-replacement techniques. Because of the inherent errors in these initial phases, subsequent phase refinement and phase extension to higher resolution can assist greatly in the early stages of structure determination. Phases can be improved and extended by density modification, which imposes physical and chemical constraints such as solvent flatness, map continuity and non-crystallographic symmetry in real space (for a review, see Podjarny, Bhat & Zwick, 1987).

Wang's (1985) solvent-flattening algorithm reduces the noise present in the diffraction data by imposing the constraint of solvent flatness. This constraint is justified because the density in the solvent region is relatively featureless compared to that of the macromolecule (Jiang & Brünger, 1994). Phases are further improved by eliminating negative densities (Schevitz, Podjarny, Zwick, Hughes & Sigler, 1981) which can arise due to series truncation errors and phase inaccuracies. Solvent flattening has met with much success and is now routinely performed, often in combination with molecular averaging (Rossmann & Blow, 1963), histogram matching (Zhang & Main, 1990), Sayre's equation (Zhang, 1993) and maximum-entropy methods (Xiang, Carter, Bricogne & Gilmore, 1993).

Solvent flattening lacks an objective criterion to assess its success. Often the performance of the method is judged by visual inspection of the resulting maps. This can be rather subjective and can lead to misassignment of protein and solvent densities in the maps. In this paper, we describe how cross-validation (Brünger, 1992*a*) can be used for assessing the quality of the solvent-flattening algorithm. Complete cross-validation (Jiang & Brünger, 1994) can be used in order to reduce fluctuations of the free R value. We show that the free R value has a high correlation with the errors in the solvent-flattened phases.

2. Materials and methods

2.1. *Test cases*

Two known protein crystal structures were chosen to assess the correlation between the free R value and the phase error with respect to the crystal structure. The test cases are examples of excellent and poor starting multiple isomorphous replacement (MIR) phases and also of complete and incomplete data.

2.1.1. *Penicillopepsin.* The first test case (an example of excellent MIR phases) was the crystal structure of penicillopepsin from *Penicillium janthinellum* consisting of 323 amino acids and 320 ordered water molecules. The space group is $C2$ with unit-cell dimensions a = 97.37, b = 46.64, c = 65.47 Å, β = 115.4° and the solvent content is 38%. It was solved by James & Sielecki (1983) with diffraction data collected at room temperature to 1.8 Å resolution by Hsu, Delbare, James & Hofmann (1977). Experimental phases to 2.8 Å were obtained from MIR using eight heavy-atom derivatives with a mean figure of merit of 0.9. The measured diffraction

ISSN 0907-4449

intensities and MIR phases are 97 and 91% complete to this resolution, respectively.

2.1.2. *Amylase inhibitor.* The second test case is an example of substantially poorer MIR phases and relatively incomplete data. The crystal structure of the α-amylase inhibitor 1HOE-467A, a small protein of 74 amino acids, was solved by Pflugrath, Wiegand, Huber & Vértesey (1986). It crystallizes in space group $P2_12_12_1$ with unit-cell dimensions a = 61.76, b = 40.73, c = 26.74 Å and has a solvent content of 50%. Diffraction data were collected at room temperature to 1.9 Å resolution and MIR phases (mean figure of merit 0.63) were obtained to 2.5 Å. The measured diffraction intensities and MIR phases are only 69.7 and 60% complete to this resolution, respectively.

2.2. *Computations*

All calculations were carried out with a developmental version of *X-PLOR* (Brünger, 1992*b*). The solvent-flattening algorithm was implemented in a major extension of the *X-PLOR* language, rather than being coded in Fortran. Excerpts of this new language and the implementation of the algorithm are described in the *Appendix*. The map calculations and Fourier transformations were performed using a grid size of 1/3 of the high-resolution limit in order to reduce problems that may arise due to undersampling.

3. Theory

3.1. *Solvent-flattening procedure*

Solvent flattening is a process that iterates between density modification in real space and phase combination in reciprocal space. The density in the putative solvent regions is replaced by, or flattened to, its average value. Positivity is enforced in the macromolecule region by the truncation of negative electron densities. Phases obtained by inverse Fourier transformation of this flattened and truncated map are then recombined with the initial phase information to produce a less biased phase estimate. These steps are summarized in Fig. 1 and described in detail in the following sections.

3.2. *Envelope calculation*

Solvent flattening requires the definition of a molecular envelope, that is, a boundary that separates solvent from macromolecule. The methods of Wang (1985) and Leslie (1988*a*) determine this boundary approximately from an initial electron density map obtained by MIR or molecular replacement.

The initial map is first truncated by setting all density points below the average to the average density. The map is then smoothed by replacing the density at each gridpoint in the initial map by the weighted mean of the electron density at all surrounding grid points within a sphere of specified radius, r_s. As the macromolecular

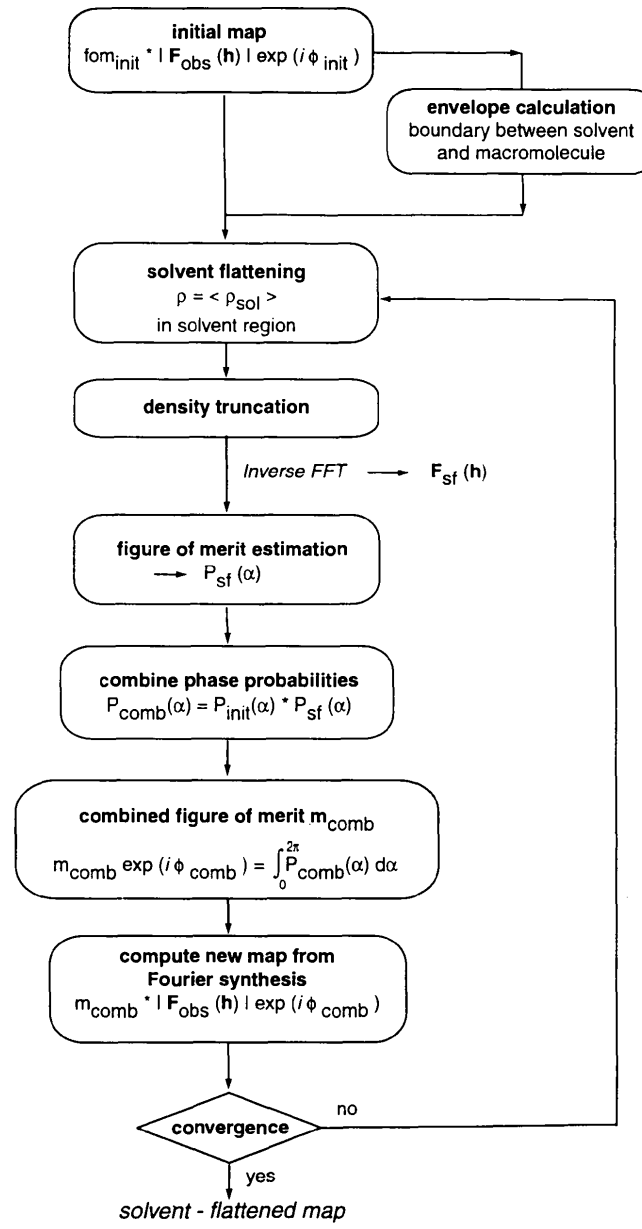

Fig. 1. Flow chart of Wang's (1985) solvent-flattening procedure as employed in this work. The molecular envelope is computed from the smoothed and truncated initial map using the method of Leslie (1988*a*). The constraints of solvent flatness and positivity are applied respectively to the solvent and macromolecule regions of the map. Updated structure factors are obtained by inverse Fourier transformation of this modified map. A figure of merit is computed for each reflection measuring the discrepancy between the original and modified structure-factor amplitudes. Phase probabilities for the modified structure factors are then computed for each reflection using (7). The modified and initial phase-probability distributions are then combined. Integration of the combined normalized probability over phase space produces a figure of merit $|\mathbf{m}_{\rm comb}|$ for each reflection. The combined phase, $\varphi_{\rm comb}$, corresponds to the new best (centroid) phase. A map is made from the new Fourier synthesis $m_{\rm comb}|\mathbf{F}_{\rm obs}(\mathbf{h})|\exp(i\varphi_{\rm comb})$ which is used for the next iteration of the whole process until the desired convergence is achieved.

densities are assumed to be higher than the solvent densities, a statistical analysis of the averaged densities can be used to discriminate between the macromolecule and solvent regions of the map. The map density points are sorted by value and a histogram of density values obtained. At this point, the solvent content of the crystal must be known which can be obtained from a crystal density experiment (Matthews, 1968). An electron-density level is determined such that the ratio of the integrals over the two portions of the histogram are equal to the solvent/macromolecule fraction. A mask is defined depending on whether the density points are below (solvent) or above (macromolecule) this density level.

The fluctuations in density in the macromolecular region, which would cause some parts of the macromolecule to appear as solvent if the initial map was used, are smeared out as a result of truncation and smoothing. Smoothing the truncated map in real space is computationally expensive, but it can be calculated more efficiently in reciprocal space (Leslie, 1988*a*). The Fourier transform of the smoothed truncated map is given by the product of the Fourier transform of the truncated initial map and an appropriate weighting function which corresponds to the smoothing operation. It is important that this double Fourier transform procedure uses all theoretically observable reflections to a certain resolution, not just the ones actually observed.

An alternative method for computing the molecular envelope consists of making a histogram of the local variation in r.m.s. densities at each gridpoint to distinguish between macromolecule and solvent (Abrahams, Leslie, Lutter & Walker, 1994). This approach assumes that the local r.m.s. variations in density are lower for solvent than for the macromolecule.

It should be noted that the accuracy of the molecular envelope critically depends on the quality of the initial phases. It is also possible to use crude initial models or to edit the mask (Jones & Kjeldgåard, 1993) before applying solvent flattening, especially when combined with NCS averaging.

Unless otherwise stated, the envelope calculations in this paper were performed only once, prior to the solvent-flattening procedure, using the Wang–Leslie method with a smoothing radius, r_s of 8 Å.

3.3. *Figure of merit estimation*

A figure of merit for the solvent-flattened phases is obtained from the dicrepancy between observed and modified structure-factor amplitudes. In the presence of an appropriate model, the figure of merit (FOM) for each phase can be estimated from,

$$\text{FOM} = \begin{cases} \tanh\left(\frac{X}{2}\right) & \text{(centrics)} \\ \frac{I1(X)}{I0(X)} & \text{(acentrics)}, \end{cases} \quad (1)$$

where $I0(X)$ and $I1(X)$ are the zero and first order modified Bessel functions of the first kind and X is given by,

$$X = 2[|\mathbf{F}_{\text{obs}}(\mathbf{h})||\mathbf{F}_{\text{calc}}(\mathbf{h})|]/\varepsilon \textstyle\sum_{\mathbf{Q}}, \quad (2)$$

$|\mathbf{F}_{\text{calc}}(\mathbf{h})|$ is the structure factor of the model, $\sum_{\mathbf{Q}}$ represents the amount of missing information and ε corrects for the difference in expected intensities for different reciprocal lattice zones (Woolfson 1956; Srinivasan & Parthasarathy 1976).

In the context of solvent flattening, the model is approximated through the solvent-flattened and truncated map, and thus $|\mathbf{F}_{\text{calc}}(\mathbf{h})|$ is set to $|\mathbf{F}_{\text{sf}}(\mathbf{h})|$ (Fig. 1). Sim (1959) suggested computing $\varepsilon \sum_{\mathbf{Q}}$ by,

$$\varepsilon\textstyle\sum_{\mathbf{Q}} = \langle|\mathbf{F}_{\text{obs}}(\mathbf{h})| - |\mathbf{F}_{\text{calc}}(\mathbf{h})|\rangle^2, \quad (3)$$

where the angular brackets $\langle\rangle$ refer to averaging over a number of resolution shells whereas Bricogne (1976) instead suggested using,

$$\varepsilon\textstyle\sum_{\mathbf{Q}} = \langle|\mathbf{F}_{\text{obs}}(\mathbf{h})|^2 - |\mathbf{F}_{\text{calc}}(\mathbf{h})|^2\rangle, \quad (4)$$

which measures the mean discrepancy between observed and calculated intensities.

Rayment (1983) proposed a much simpler figure of merit estimation in phase refinements of the structure of mouse polyoma virus capsids,

$$\text{FOM} = \exp\left[-\left||\mathbf{F}_{\text{obs}}(\mathbf{h})| - |\mathbf{F}_{\text{calc}}(\mathbf{h})|\right|/|\mathbf{F}_{\text{obs}}(\mathbf{h})|\right]. \quad (5)$$

Little theoretical justification for any of these figure of merit estimations exists in the context of solvent flattening. The performance of all three methods is empirically assessed in §4.

3.4. *Phase combination*

Solvent flattening is prone to model bias when errors are present in the solvent/molecule boundary. Consequently, portions of molecular density can be flattened by uncritical application of this method. Model bias can be reduced by combining phase probability distributions of solvent-flattened phases with those of the initial phases.

Phase combination is achieved by multiplication of the phase probabilities,

$$P_{\text{comb}}(\alpha) = [P_{\text{init}}(\alpha)][P_{\text{sf}}(\alpha)], \quad (6)$$

where $P_{\text{init}}(\alpha)$ is the phase probability distribution of the initial phases and $P_{\text{sf}}(\alpha)$ is that derived from the solvent-flattened density map.

By equating the model with the truncated and solvent flattened map, the following phase probability distribution (Hendrickson & Lattman, 1970; Table 9.1 of Srinivasan & Parthasarathy, 1976) is appropriate,

$$P_{sf}(\alpha) = \begin{cases} N\exp[(X/2)\cos(\varphi_{calc})\cos\alpha + (X/2)\sin(\varphi_{calc})\sin\alpha] & \text{(centrics)} \\ N\exp[X\cos(\varphi_{calc})\cos\alpha + X\sin(\varphi_{calc})\sin\alpha] & \text{(acentrics)} \end{cases}. \quad (7)$$

On integration of the normalized combined probability distribution, $P_{comb}(\alpha)$, over phase space, a combined best (centroid) phase and figure of merit can be obtained for each reflection,

$$m_{comb}|\mathbf{F}_{obs}(\mathbf{h})|\exp(i\varphi_{comb}) = 1/N \int_0^{2\pi} |\mathbf{F}_{obs}(\mathbf{h})| \times [P_{comb}(\alpha)dx], \quad (8)$$

where N is a normalization factor. The left-hand side of (8) is the most common Fourier synthesis used for solvent flattening. Alternative forms of Fourier syntheses are briefly discussed in §4.3.

3.5. *Density truncation*

The F_{000} term is in general unknown for macromolecular crystal structures. Thus, the average of the electron-density map is zero which will affect the result of density truncation. An estimate for the average electron density (F_{000}/V) (Leslie, 1988*b*) can be obtained from,

$$[(F_{000}/V) + \langle\rho_{sol}\rangle]/[(F_{000})/V) + \langle\rho_{macro}\rangle] = S, \quad (9)$$

where $\langle\rho_{sol}\rangle$ and $\langle\rho_{macro}\rangle$ are the average solvent and macromolecular densities in the Fourier map (computed excluding the F_{000}/V term). S is the ratio of the physical solvent and macromolecular densities. For proteins, the density is around 0.43 e^- Å^{-3}. The solvent density is dependent on the environment. For water this is 0.33 e^- Å^{-3} leading to a value of S of 0.77. However, this is only appropriate if *all* of the low resolution terms are present. Thus, to obtain maximum performance of the algorithm it may be advisable to treat S as an adjustable parameter.

The average electron density F_{000}/V is added to the map prior to truncation of negative densities.

3.6. *Measure of convergence*

The convergence of the solvent-flattening algorithm can be assessed by the R value between observed and modified structure factors [$\mathbf{F}_{obs}(\mathbf{h})$ and $\mathbf{F}_{sf}(\mathbf{h})$] weighted by the combined figure of merit m_{comb}.

$$R = \sum_{\mathbf{h}} m_{comb}(\mathbf{h})\big||\mathbf{F}_{obs}(\mathbf{h})| - k|\mathbf{F}_{sf}(\mathbf{h})|\big| \div \sum_{\mathbf{h}} m_{comb}(\mathbf{h})|\mathbf{F}_{obs}(\mathbf{h})|, \quad (10)$$

where k is a scale factor. We used $m_{comb}(\mathbf{h})$ weighting in calculating the R value because it is supposed to assess the quality of the electron-density maps computed using figure-of-merit weighted amplitudes (8).

3.7. *Complete cross validation*

The R value (10) is a poor criterion for assessing phase accuracy because it can be made arbitrarily small by inappropriate paramaterization, *e.g.* by using a very small smoothing radius. As a result, the data are overfit, producing possibly poorer phases than the initial ones. This problem can be avoided by using cross validation. The free R value shows a much higher correlation with the phase accuracy of refined models than the R value (Brünger, 1992*a*, 1993). This method has already been used to optimize the performance of density skeletonization (Baker, Bystroff, Fletterick & Agard, 1993; Grimes & Stuart, 1994).

In the context of solvent flattening, cross-validation consists of omitting a certain subset or test set, T, of the observed data, modifying electron-density maps computed using the remaining data, updating calculated structure factors [$F_{sf}(\mathbf{h})$] by inverse Fourier transformation, and evaluating the free R value (R_{free}) over the test set of reflections T,

$$R_{free} = \sum_{\mathbf{h}\in\mathbf{T}} m_{comb}\big||\mathbf{F}_{obs}(\mathbf{h})| - k|\mathbf{F}_{sf}(\mathbf{h})|\big| \div \sum_{\mathbf{h}\in\mathbf{T}} m_{comb}|\mathbf{F}_{obs}(\mathbf{h})|. \quad (11)$$

The choice of the test set usually has little influence on the behavior of the free R value, provided the selection is purely random and the test set contains a sufficient number of data points (Brünger, 1993). However, considerable variation of R_{free} can occur at low resolution because there are comparatively few reflections (Jiang & Brünger, 1994). In crystallographic refinement the low resolution reflections are usually omitted, but they are of vital importance for solvent flattening since they affect the definition of the envelope and the connectivity of the flattened map. To reduce fluctuations of the free R value at low resolution one can use complete cross-validation (Jiang & Brünger, 1994). Briefly, the observed diffraction data set is partitioned into n non-overlapping test sets ($T_1, \ldots, T_n$) where each set contains a subset (*e.g.* 10%) of the data. For each test set T_i, a corresponding working set A_i is defined consisting of all data excluding T_i. Solvent flattening is carried out n times, once for each of the working sets A_i. Only the working reflections are used to compute the electron-density map, but updated solvent-flattened calculated structure factors are obtained for all the reflections, including the test set T_i, on inverse Fourier transformation of the modified map to reciprocal

space (Fig. 2). After the n separate solvent-flattening macrocycles have been performed, the structure factors $F_{cv}^{sf}(\mathbf{h})$ for the test sets are merged and the completely cross-validated R value computed,

$$R_{free}^{complete} = \sum_{\mathbf{h}} m_{comb} \big| |\mathbf{F}_{obs}(\mathbf{h})| - k|\mathbf{F}_{cv}^{sf}(\mathbf{h})| \big| \div \sum_{\mathbf{h}} m_{comb} |\mathbf{F}_{obs}(\mathbf{h})|. \quad (12)$$

It should be noted both the working R value (10) and the free R values for an individual test set (11) and the completely cross-validated R value (12) are zero prior to density modification. Thus, the performance of solvent flattening is monitored by the change in the R value and the free R values after the first density-modification cycle.

Cross-validation is useful for testing and optimization of the algorithm.

3.8. *Phase refinement versus phase extension*

Solvent flattening can be used to extend phases to higher resolution as well as to refine them. Since initial phases are unavailable for those that are extended the algorithm described above must be slightly modified. Phase combination cannot be performed for the phase-extended reflections. Thus, the phase probability distribution from solvent flattening, $P_{sf}(\alpha)$, is directly used. Only the initially phased reflections are used to compute the initial map. In subsequent cycles all observed diffraction data are included, thus producing phases for initially unphased reflections. The figure of merit estimations (1)–(5) are computed for all observed intensities.

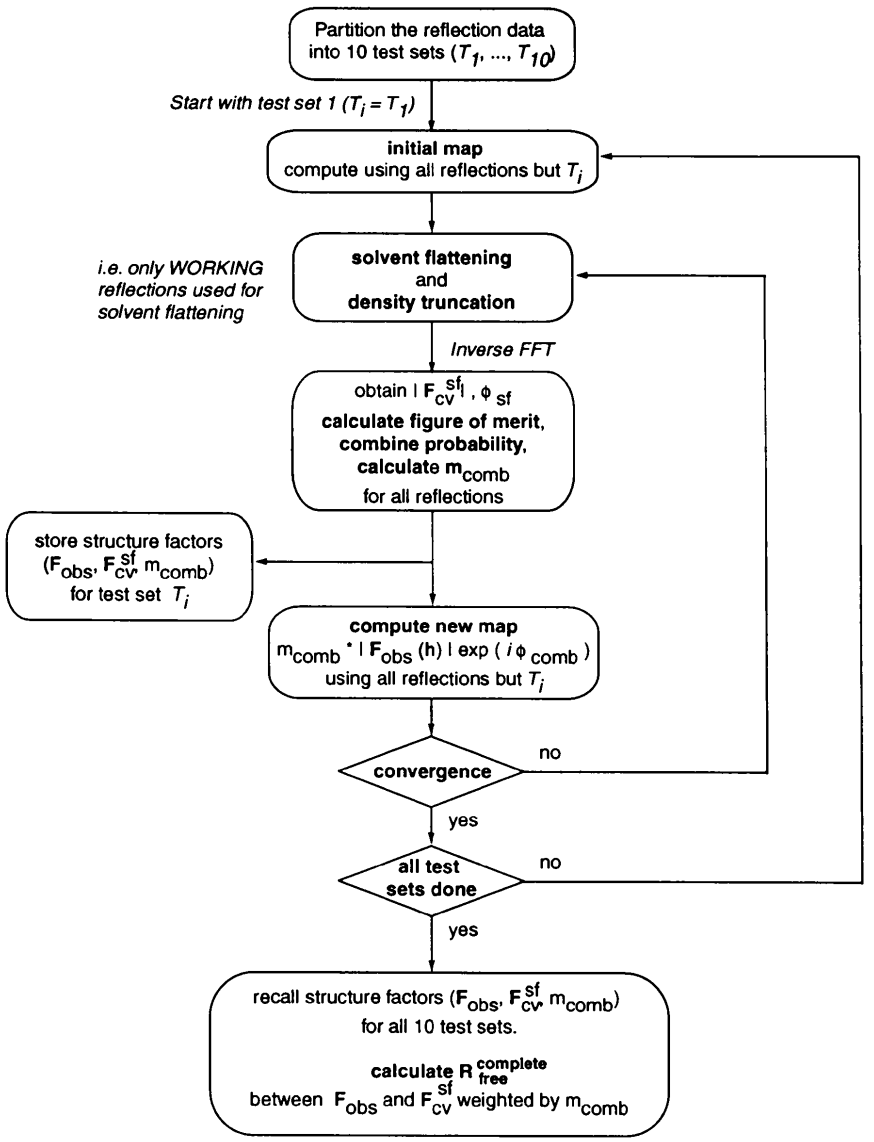

Fig. 2. Flow chart of complete cross-validation applied to solvent flattening. The molecular envelope is computed using all observed reflections prior to the cross-validated solvent-flattening process. Only the working reflections (all but T_i) are used to compute the electron-density maps. The inverse Fourier transformation updates calculated structure factors, $\mathbf{F}_{sf}(\mathbf{h})$, for *all* observed reflections. Structure factors for the current test set, $F_{cv}^{sf}(\mathbf{h})$, are written out for each cycle of the refinement. When all ten refinements are completed these structure factors are used to compute $R_{free}^{complete}$ (12).

We will mainly focus on phase refinement in this paper although cross-validation can be applied to phase extension as well. For phase refinement, the test sets are chosen from the set of all reflections with known initial phases. For phase extension, the test sets are chosen from all reflections for which intensities have been measured.

3.9. *Testing the algorithm*

For a known crystal structure, the quality of the solvent-flattened phases can be assessed by the phase difference between the combined phases, $\varphi_{\text{comb}}(\mathbf{h})$, from solvent flattening and those calculated from the model, $\varphi_{\text{model}}(\mathbf{h})$,

$$\langle m_{\text{comb}}(\mathbf{h})\Delta\varphi\rangle = \sum_{\mathbf{h}} m_{\text{comb}}(\mathbf{h})|\varphi_{\text{comb}}(\mathbf{h}) - \varphi_{\text{model}}(\mathbf{h})| \div \sum_{\mathbf{h}} m_{\text{comb}}(\mathbf{h}). \quad (13)$$

Figure-of-merit weighting is applied as in the computation of R (10), R_{free} (11) and $R_{\text{free}}^{\text{complete}}$ (12). The phase errors were always obtained for all reflections for which initial phases were available regardless of cross-validation. In a real situation, phase accuracy cannot be assessed because the crystal structure is in general unknown.

4. Results and discussion

4.1. *Influence of the completeness of the data*

The influence of the completeness of the data on the accuracy of the free R value is shown in Figs. 3 and 4 for penicillopepsin and the amylase inhibitor data, respectively. Complete cross-validation was repeated five times using different random assignments of the test sets to examine the variation of $R_{\text{free}}^{\text{complete}}$ for phase refinement. The free R value shows considerable variation when computed for the individual test sets for both penicillopepsin (~0.03) and amylase inhibitor (~0.13). To reduce these variations we used complete cross validation. The $R_{\text{free}}^{\text{complete}}$ values show smaller variations for different partitionings of the data compared to the individual free R values (Figs. 3*b* and 4*b*). The errors on the mean associated with the average value are around 0.002 and 0.003 for penicillopepsin and amylase inhibitor, respectively. This suggests that

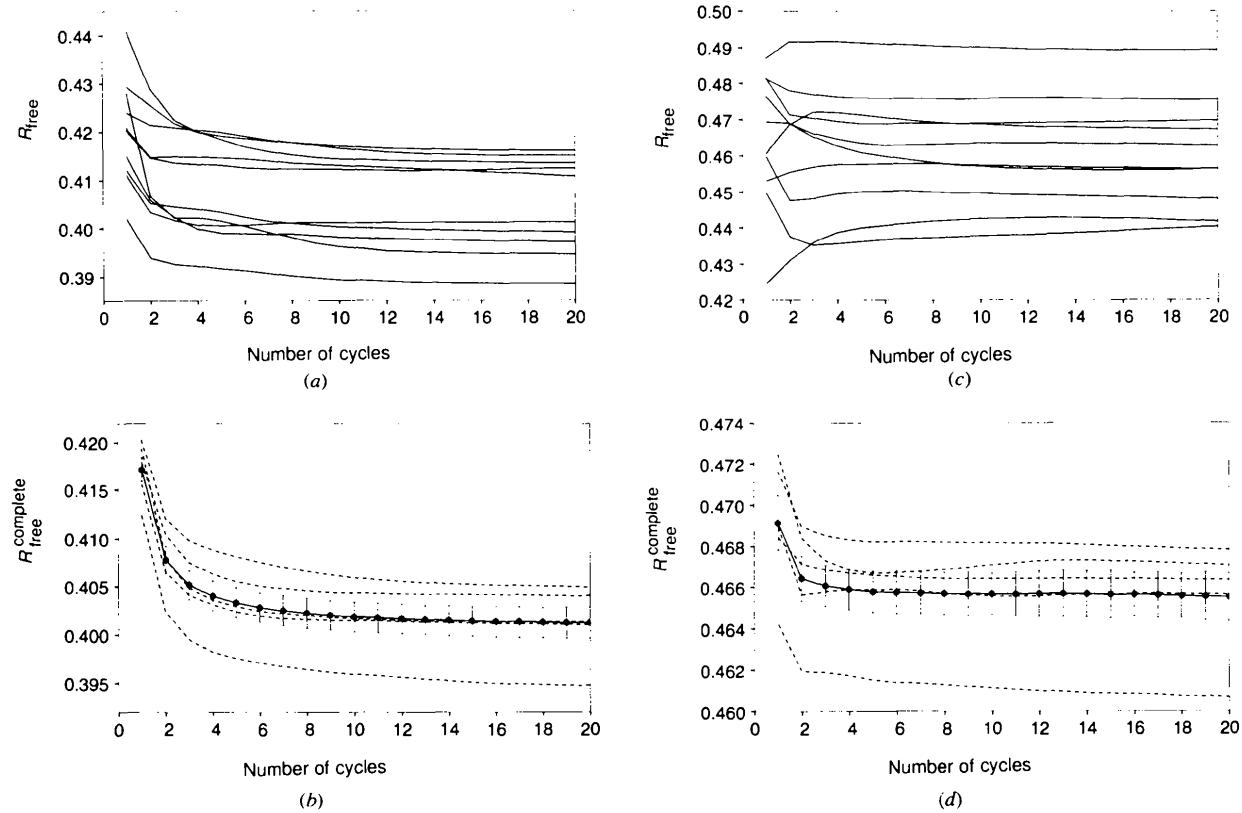

Fig. 3. The accuracy of cross-validated R values for penicillopepsin using different partitionings of the diffaction data. Phase combination was carried out using $u = 1$, $v = 1$ (14) and density truncation with $S = 0.87$ (9). (*a*) and (*b*) were generated using all initially phased reflections. (*c*) and (*d*) were obtained using only 70% of these reflections (selected randomly) to examine the success of complete cross-validation with incomplete data. The solid lines in (*a*) and (*c*) correspond to the individual free R values for the ten test sets using a random partitioning of the data set. In (*b*) and (*d*) the dashed lines show $R_{\text{free}}^{\text{complete}}$ (12) for five different random partitionings. The solid line with circles shows the mean value and the errors on the mean (defined as r.m.s.$(R_{\text{free}}^{\text{complete}})/n^{\frac{1}{2}}$ where n is the number of partitionings).

it would be ideal to repeat complete cross-validation many times in order to obtain a converged average and reduced errors of the mean. However, at least for penicillopepsin the $R_{free}^{complete}$ curves obtained from five independent partitionings are similar (Fig. 3*b*). This shows that a single complete cross-validation should be sufficient in this case.

The results for amylase inhibitor are not quite as encouraging as $R_{free}^{complete}$ shows large fluctuations for the different partitionings (Fig. 4*b*). These fluctuations arise because of both incompleteness of the data (about 60% complete) and the small overall number of the unique set of reflections.

The influence of completeness and data-set size have been studied by randomly removing 30% of the experimental data in penicillopepsin to simulate incomplete data (Figs. 3*c* and 3*d*). The effect of removing 30% of the data, at random, for penicillopepsin is to slightly increase the fluctuations among the single test sets (Fig. 3*c*). Furthermore, the mean of $R_{free}^{complete}$ shows large errors (Fig. 3*d*). In light of these results, using only a single test set to assess the quality of density modification procedures can be prone to error. Even for the almost complete data of penicillopepsin the free R value shows variations among the ten test sets with a standard deviation of around 1%, consistent with the calculations of Brünger (1995). It is, therefore, advisable to use complete cross-validation and in cases of incomplete data or small data-set sizes a mean value for completely cross-validated R values for a number of different partitionings of the data set should be obtained.

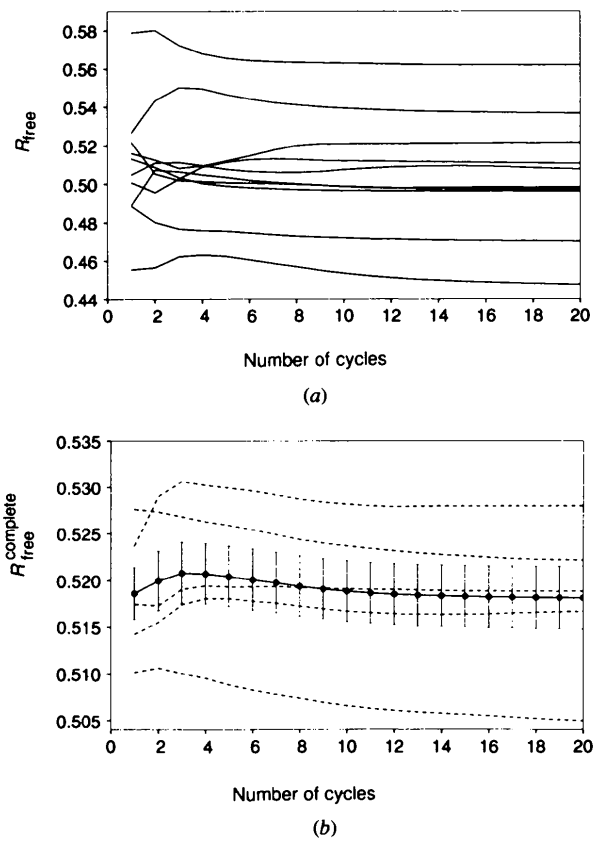

Fig. 4. The accuracy of cross-validated R values for the amylase inhibitor. Phase combination was carried out using $u = 1$, $v = 1$ (14) and density truncation with $S = 0.70$ (9). (*a*) and (*b*) were generated using all the initially phased reflections which were only 60% complete. The solid lines in (*a*) correspond to the individual free R values for the ten test sets using a random partitioning of the data set. In (*b*) the dashed lines show $R_{free}^{complete}$ (12) for five different random partitionings. The solid line with circles shows the mean value and the errors on the mean (defined as r.m.s.$(R_{free}^{complete})n^{\frac{1}{2}}$ where n is the number of partitionings).

4.2. *Figure of merit estimation*

Different figure of merit estimates [(1)–(5)] are compared in Fig. 5 for phase refinement of penicillopepsin at 2.8 Å resolution. Bricogne's modified version of Sim weighting (4) gives rise to the most stable behavior and largest improvement, as assessed by both the free R value and the phase errors. The Rayment weighting scheme (5) gives rise to slightly oscillatory phase errors and free R values. Sim weighting (3) produces the lowest working set R value but the phase error and free R value diverge as the number of cycles increases.

4.3. *Phase combination*

Equal weighting of the phase probabilities (6) does not necessarily lead to the best combined phases as suggested by Bricogne (1976). This is because the two phase probability distributions (6) are not completely independent. The solvent-flattened phase probability distribution, $P_{sf}(\alpha)$, is obtained from the Fourier transformation of the modified electron-density map which was in turn derived from the initial phases with phase probability distribution $P_{init}(\alpha)$. Bricogne (1976) concluded that a solution to this problem is a relative weighting of the two sources of phase information, as in,

$$P_{comb}(\alpha) = [P_{init}(\alpha)]^u [P_{sf}(\alpha)]^v, \qquad (14)$$

with u not equal to v. A similar issue is encountered when combining sources of phase information for different derivatives in MIR phasing (Blow & Matthews, 1973). A related approach to overcome the problems of combining partially dependent phase probabilities involves adjusting the amplitude coefficients in the resulting combined Fourier synthesis. Rice (1981) suggested using Fourier syntheses of the form,

$$\mathbf{F}_{comb}(\mathbf{h}) = m_{comb}\{|\mathbf{F}_{obs}(\mathbf{h})| + Q_c[|\mathbf{F}_{obs}(\mathbf{h})| - |\mathbf{F}_{calc}(\mathbf{h})|]\} \times \exp(i\varphi_{comb}), \qquad (15)$$

with Q_c set to 3. Stuart & Artymuik (1985) and Read (1986) extended this approach further by calculating val-

ues of Q_c for each reflection from the errors associated with each of the phase distributions.

We have used complete cross-validation to identify the optimum relative weights u and v in (14) using the standard Fourier synthesis (8). We have constrained $u + v$ to a value of two such that the combined probability distribution is not broadened or sharpened compared to the commonly used unit weighting. For penicillopepsin, significant phase improvement was gained while preserving converged behavior by using powers u and v of 0.75 and 1.25, respectively (Fig. 6*c*). The corresponding $R_{\text{free}}^{\text{complete}}$ value (Fig. 6*b*) is essentially converged and is marginally lower than for $u = 1$ and $v = 1$. However, after about ten cycles the $R_{\text{free}}^{\text{complete}}$ values for these two combination schemes are almost identical. Given the error in $R_{\text{free}}^{\text{complete}}$ of about 0.5% (Fig. 3*b*), these differences cannot be regarded as significant. Using an even lower value, $u = 0.25$, initially leads to reduced

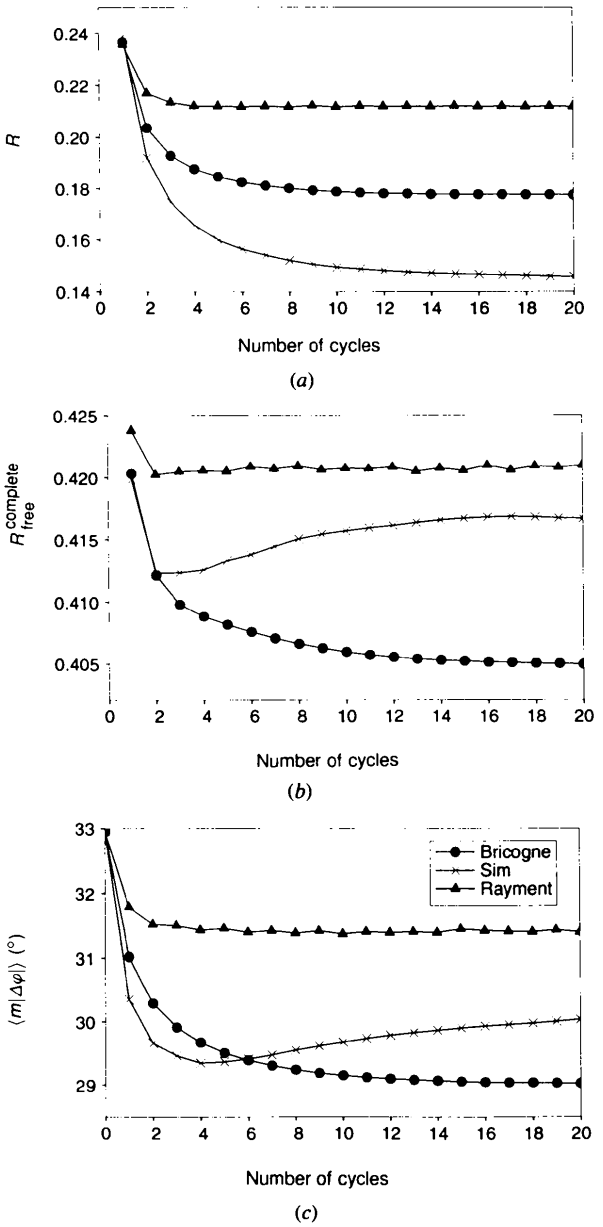

Fig. 5. Comparison of methods for estimating the figure of merit of the solvent-flattened phases for penicillopepsin. Circles: Bricogne's version of Sim weighting (4); crosses: Sim weighting (3); triangles: Rayment weighting (5). Phase combination was performed with equal weighting applied to the component phase probability distributions and the level of density truncation was set to $S = 0.87$. $R_{\text{free}}^{\text{complete}}$ was computed from one complete partitioning of the data.

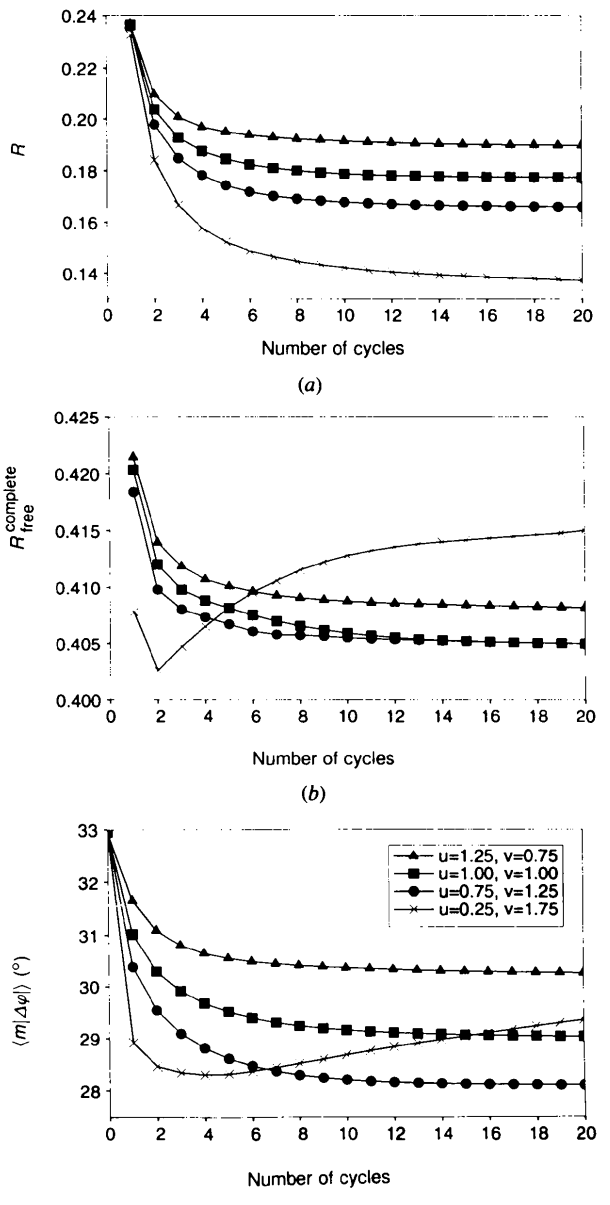

Fig. 6. Comparison of phase probability combinations for penicillopepsin. Bricogne's formula (4), $S = 0.87$. Triangles: $u = 1.25$, $v = 0.75$; squares: $u = 1$, $v = 1$; circles: $u = 0.75$, $v = 1.25$; crosses: $u = 0.25$, $v = 1.75$. $R_{\text{free}}^{\text{complete}}$ was computed from one complete partitioning of the data.

values in both the phase errors and the $R_{free}^{complete}$ value for the first few cycles. However, divergence emerges during subsequent cycles even though the conventional R value (10) continues to decrease monotonically. This behavior is indicative of overfitting the observed amplitudes by putting too little weight on the MIR phases. Complete cross validation can be used to detect the point at which the data become overfit as a result of drifting away from the initial MIR solution. On the other hand, in the case where too much weight is applied to the MIR phases (u = 1.25, v = 0.75) the phase errors, $R_{free}^{complete}$ values and R values are all higher. While the free R value is not perfectly correlated with phase accuracy, it nevertheless shows much better behaviour than the normal R value. The combination scheme u = 0.75 and v = 1.25 with the lowest R value that does not overfit the data (as detected by $R_{free}^{complete}$) is the most appropriate method as judged by the phase errors.

4.4. *Density truncation*

Complete cross-validation was used to assess the effect of density truncation (Fig. 7). Both $R_{free}^{complete}$ and the phase errors indicate that solvent flattening is significantly improved when density truncation is employed. We used a value of S = 0.87. Slightly higher or lower values of S (for example, S = 0.60) also give rise to converged $R_{free}^{complete}$ values and phase errors, but the performance is slightly poorer. However, these differences in $R_{free}^{complete}$ are often insignificant given the error of around 0.5% as estimated from Fig. 3(*b*).

4.5. *Envelope calculations*

The choice of the smoothing radius used for the envelope calculation has been the subject of much discussion by Leslie (1988*b*). Too small a radius is likely to produce cavities within the protein and assign external protein loops to the solvent region, whereas too large a value will remove detail from the solvent/macromolecule boundary. This is reflected in the poor behaviour of $R_{free}^{complete}$ and the phases error for small (2 Å) and large (15 Å) values of the smoothing radius (Fig. 8). In contrast, the normal R value is best for the largest phase errors (2 Å).

4.6. *Phase extension*

Solvent flattening was used to extend the resolution of the initial phases (Fig. 9). Complete cross-validation showed that optimal parameters for phase extension were close to those of phase refinement (combination scheme u = 0.75, v = 1.25 with Bricogne's version of Sim weighting and density truncation applied with S = 0.87). A fairly low phase error was achieved for the extended reflections illustrating the power of phase extension using solvent flattening; the constraint of flat solvent density improves the phases at high resolution because it removes some of high resolution noise present in the data. The performance of the phase improvement method is remarkable considering that phase extension was performed in one step from 2.8 to 1.8 Å resolution. Usually, phase extension is performed in several steps in reciprocal space. Clearly, this unexpected observation needs further investigation.

4.7. *Other methods*

Phase combination of the modified phases with the initial experimental phases helps eliminate model bias. However, if some *noisy* density in part of the solvent

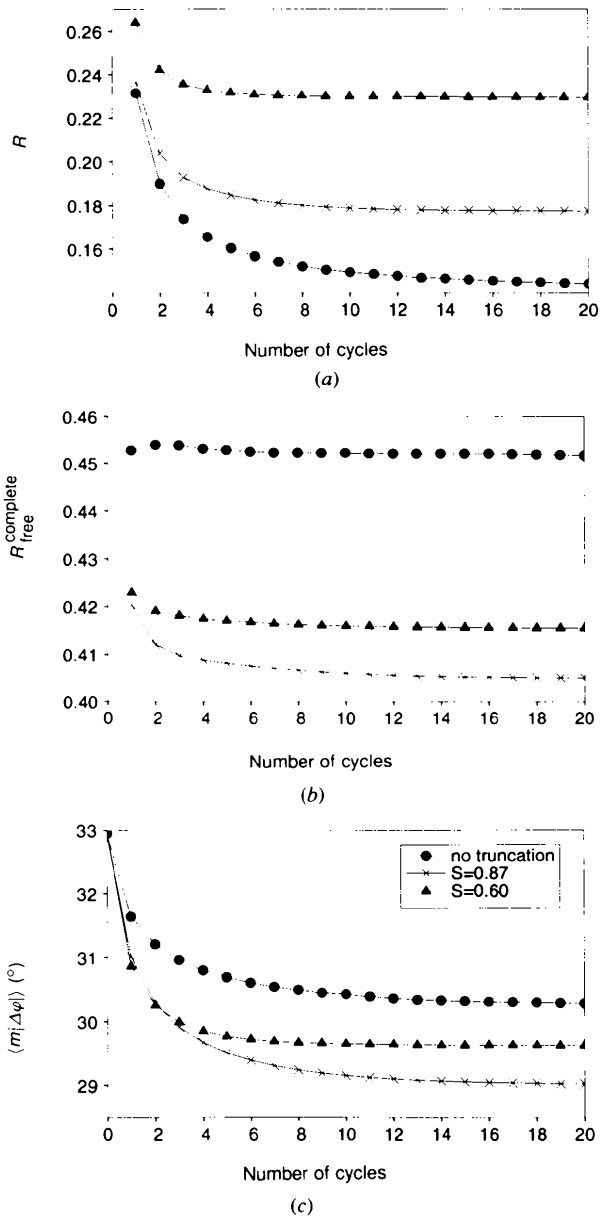

Fig. 7. Effect of truncation on solvent-flattened phases for penicillopepsin. Bricogne's formula (4), u = 1, v = 1. Circles: no truncation; crosses: truncation with S = 0.87; triangles: truncation with S = 0.60. $R_{free}^{complete}$ was computed from one complete partitioning of the data.

region is removed by flattening then this original, undesired, feature will to some extent still be present in the resulting *combined* Fourier synthesis. This may lead to slower convergence of the algorithm. Abrahams *et al.* (1994) proposed a possible solution to this problem by *flipping* rather than flattening the solvent densities. In this manner, originally more negative regions of electron density become positive on flipping such that the combined phases are more likely to lead to a *flat* region in the resulting electron-density map.

We find that flipping rather than flattening the solvent densities and computing the molecular envelope by local r.m.s. variations in density (see §3.2) yields only slight improvement for the penicillopepsin test case (not shown). Only marginal or no improvement is achieved using these methods possibly as a result of the high quality of the initial MIR phases. However, in the case of the F_1-ATPase structure solved by Abrahams *et al.* (1994), which had very much poorer starting experimental phases, significant phase improvement was obtained using both methods.

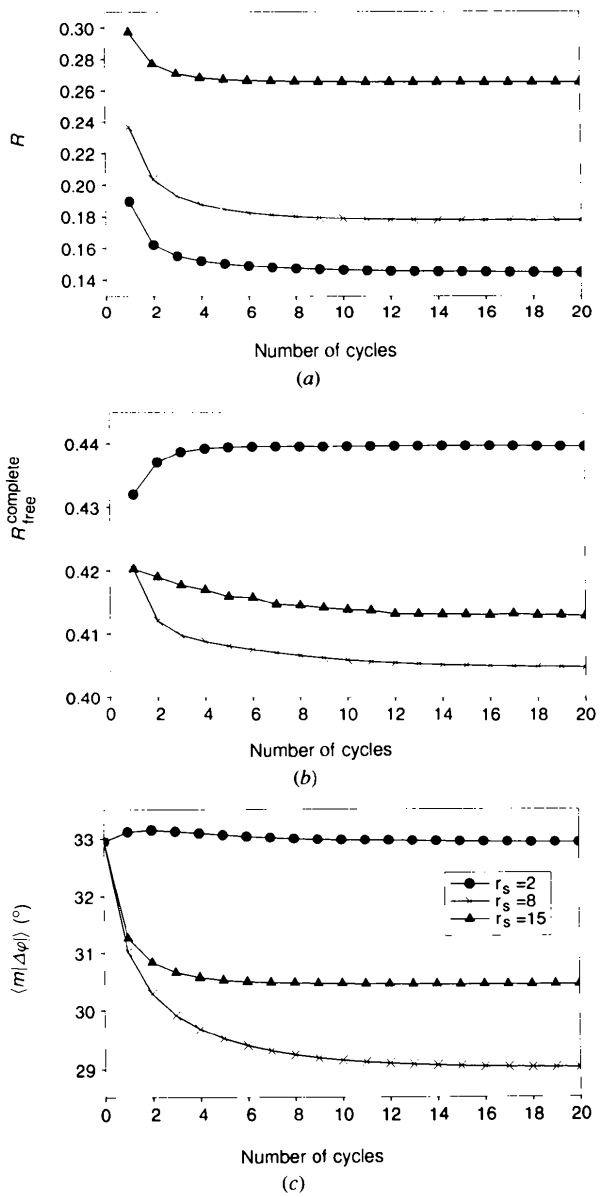

Fig. 8. Comparison of different smoothing radii r_s. Bricogne's formula (4), S = 0.87, u = 1, v = 1. Circles: r_s = 2 Å; crosses: r_s = 8 Å; triangles: r_s = 15 Å. $R_{free}^{complete}$ was computed from one complete partitioning of the data.

5. Concluding remarks

Solvent flattening can be a useful tool for improving experimental or molecular-replacement phases. However, up to now no objective criterion has been available by which to assess and quantify the improvement. We have shown that cross-validation is a step towards developing such a criterion. There was a high correlation between the free R value and phase errors (Figs. 5–8) except for the phase combination scheme u = 0.75, v = 1.25 (Fig. 6). This degree of correlation suggests that complete cross-validation can be used to identify cases where solvent flattening significantly overfits the data or where the improvement is minimal.

The free R value obtained from solvent flattening is fairly sensitive to the partitioning of the data into test and working sets. This is especially true if the number of reflections is small due to incompleteness of the data or small unit cell size. This problem is largely overcome by using complete cross-validation where the reflection data are partitioned into ten non-overlapping test sets. $R_{free}^{complete}$ is obtained after running cross-validation ten times, once for each of the test sets.

Some phase improvement is achieved when no density truncation is applied (Fig. 7). However, both the free R value and the phase errors decrease significantly

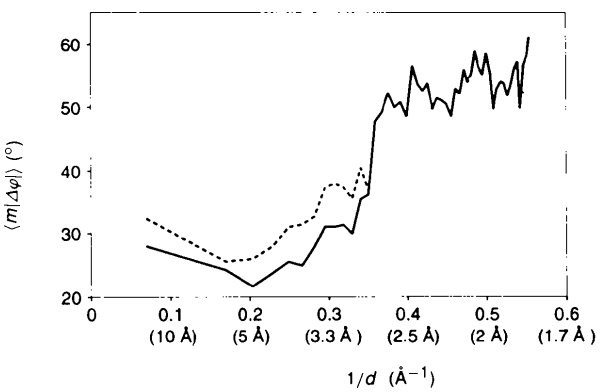

Fig. 9. Phase extension for penicillopepsin to 1.8 Å resolution. Shown are the figure of merit weighted phase errors (13) both before (dashed line) and after (solid line) 20 cycles of solvent flattening. r_s = 8 Å, S = 0.87, Bricogne's formula (4) and phase combination procedure u = 0.75, v = 1.25. $1/d$ greater than 0.357 Å^{-1} corresponds to the phase-extended region.

when negative densities are eliminated. On the other hand, the working R value is considerably lower when no truncation is enforced making it an unsatisfactory indicator of phase improvement. Thus, the free R value is successful in identifying the method which gives the better phase improvement.

In a similar vein, overfitting occurs when the smoothing radius of the Wang–Leslie envelope mask is too small (Fig. 8). The free R value and the phase errors are signficantly higher with an averaging radius of 2 Å rather than 8 Å while the conventional R value is lower. Thus, the free R value is correlated with phase accuracy whereas the R value is often not correlated.

We chose as our primary test case the almost complete diffraction data of penicillopepsin with very good quality starting experimental phases. Although this structure is not necessarily representative of a typical case for solvent flattening, the main point of this paper has been to demonstrate the validity of cross-validation for monitoring density-modification methods such as solvent flattening. Optimal parameters have been derived for solvent flattening of penicillopepsin, but it should be noted that these parameters do not necessarily hold for other systems. Indeed, the case of using the methods of inverting the solvent densities and calculating the molecular envelope *via* local r.m.s. variations in density (see §4.7) serves as an example; no further improvements over conventional methods are achieved with penicillopepsin whereas dramatic phase improvements are reported using these new methods for the *F*1 structure (Abrahams *et al.*, 1994). Thus, cross validation should be used to optimize the performance of density modification on a case-by-case basis.

We thank J. S. Jiang, P. Gros and W. L. DeLano for contributions to this work and E. M. Duffy, A. M. Friedman, V. L. Rath and L. M. Rice for critically reading the manuscript. This work was supported by a grant from the NSF to ATB (DIR 9021975). ALUR acknowledges support from a NIH grant (GM 22778-18) to Hal Wyckoff prior to this project.

APPENDIX

A1. Overview of X-PLOR's crystallographic language

Solvent flattening has been implemented using a new crystallographic language developed in *X-PLOR*. This language consists of concise structure-factor and map manipulations. The key features of this new language include:

- the ability to declare arrays in real and reciprocal space;
- the use of selections for operations involving both structure factors and map elements;
- modularity as accomplished through macros;
- the possibility to design new algorithms without having to alter the underlying Fortran code.

A2. Solvent-flattening algorithm

An extract from the main body of the *X-PLOR* script for phase refinement is shown below.

```
{* PHASE IMPROVEMENT CYCLES *}

while ($1 < $ncycles) loop main
   evaluate ($1=$1+1)

  {* DENSITY MODIFICATION *}
  {* apply solvent flattening *}
  do (map=$ave_sol_den) (mask=1)

  {* apply negative density elimination *}
  do (map=-$f000_over_v) (map < -$f000_over_v)

  {* inverse FFT to update structure factors Fsf *}
  do (fsf=ft(map)) (all)

  {* COMPUTE NEW FIGURE OF MERIT *}
  do (fomc=0) (all)

  @MACRO:weighting ( #f1=fobs;
                     #f2=fsf;
                     #fom=fomc;
                     #sel=(amplitude(fobs)>0 and fom>0); )

  {* truncate figure of merit to avoid overflow problems *}
  do (fomc=min(fomc,0.9995)) (all)

  {* calculate Hendrickson-Lattman coefficients from the *}
  {* modified phases and fom (fomc) *}
  @MACRO:fomtox ( #xfit=xc;
                  #fom=fomc; )

  do (HLAA =xc*cos(phase(fsf))) (all)
  do (HLBB =xc*sin(phase(fsf))) (all)
  do (HLCC =0) (all)
  do (HLDD =0) (all)

  {* PHASE COMBINATION *}
  {* first reset initial Hendrickson-Lattman coefficients *}
  do (HLA =xo*cos(phase(fobs))) (all)
  do (HLB =xo*sin(phase(fobs))) (all)
  do (HLC =0) (all)
  do (HLD =0) (all)

  {* combine probabilities *}
  @MACRO:combineprobability ( #messages="off";
                              #addname="from initial phases";
                              #pa=HLA;
                              #pb=HLB;
                              #pc=HLC;
                              #pd=HLD;
                              #w=$u;
                              #addname="from density modification";
                              #adda=HLAA;
                              #addb=HLBB;
                              #addc=HLCC;
                              #addd=HLDD;
                              #addw=$v;)

  {* use only phase probabilities from density modification for *}
  {* initially unphased reflections *}
  do (HLA =HLAA) (fom <=0)
  do (HLB =HLBB) (fom <=0)
  do (HLC =0) (fom <=0)
  do (HLD =0) (fom <=0)

  {* compute the combined figure of merit * }
  @MACRO:getfom ( #pa=HLA;
                  #pb=HLB;
                  #pc=HLC;
                  #pd=HLD;
                  #m=m; )
  {* compute new Fourier synthesis from combined fom and combined phase *}
  do (fcomb=combine(amplitude(m)*amplitude(fobs),phase(m))) (all)

  {* prepare for next cycle - make map from new Fourier synthesis*}
  do (map=ft(fcomb)) (amplitude(fobs)> 0 and fom>0)

  {* OUTPUT STATISTICS *}

  set display=$output_Rvalue end
  display cycle $1
  @MACRO:statistics_Rvalue ( #fobs=fobs;
                             #fcalc=fsf;
                             #fom=fom;
                             #ampm=amplitude(m);)

  set display=$display_Rvalue end
  display $1[f3.0] $e2i[f6.4]

end loop main
```

Objects in real space are:

`map` – the electron-density map, subject to density modification;

`mask` – the solvent/macromolecule mask map;

and in reciprocal space:

`fsf` – modified structure factors prior to phase combination;

`fcomb` – phase combined structure factors;

`fomo` – figure of merit of the initial phases;

`fomc` – figure of merit of the calulated phases from density modification;

`m` – complex figure of merit of the combined phases [where `amplitude(m)` is the actual figure of merit; and `phase(m)` the best combined (centroid) phase)] `xo` and `xc` represent the initial and calculated values of `X` (2) required for computation of the phase probability distribution (7);

`HLA` *etc.* and `HLAA` *etc.* are the Hendrickson–Lattman coefficients (Hendrickson & Lattman, 1970) of the initial and calculated phase probability distributions. The combined probability is returned as `HLA` *etc.*

The F_{000}/V term (`$f000_over_v`) and the average solvent density (`$ave_sol_den`) in the map, $\langle\rho_{sol}\rangle$, are calculated prior to the phase improvement cycles. The relative weights for phase combination, `$u` and `$v`, are set at the beginning of the script.

The changes that had to be made to the script for cross-validation were minimal as a result of the selection operator. Maps for density modification can be made using only the working reflections by including the additional selection (`test = 0`), *i.e.* by changing the selections from `(amplitude(fobs)> 0 and fom > 0)` to `(amplitude(fobs) > 0 and fom > 0 and test = 0)`.

The selection facility allows a high degree of control over which elements are selected for a particular operation. A full aymmetric unit of reflections have been defined for the computation of the Wang–Leslie mask. Thus, it is necessary to use the selection `(amplitude(fobs) > 0)` to prevent unobserved reflections with `fobs = 0` from being included in the shell-wise average of $\varepsilon\sum_{\mathbf{Q}}$. In the case of refinement it is also necessary to use the additional selection `(fom > 0)` which flags the initially phased reflections.

*A*3. *Macros*

An *X-PLOR* macro is a script file with a well defined set of input parameters (W. L. DeLano, unpublished results). These macro parameters are declared in a header at the top of the script along with optional default values. When a macro is invoked from *X-PLOR*, the parameters and any default values are read from the macro file header. These parameters can then be modified upon macro invocation. All other parameters are assigned different values before the body is evaluated. The following example is the macro file used to calculate the Wang–Leslie envelope mask.

```
macro {wang_leslie_mask}
(
   #map1=map;
   #map2=mask;
)

   declare name=wl domain=reci type=real end  {* wl is the weighting function*}
                                              {* in reciprocal space        *}
   declare name=temp domain=reci type=comp end        {* temporary structure *}
                                                      {* factor array        *}

   {* compute weighting function wl *}
   evaluate ($rtod=180.0/$pi)
   do (wl=2.0*$pi*s()*$Rs ) (all)
   do (wl=3(sin(wl*$rtod)-wl cos(wl*$rtod)) / wl^3
           - 3(2 wl sin(wl*$rtod)
           - ( wl^2-2) cos(wl*$rtod) - 2) / wl^4) (all)

   do (temp=wl * ft(max(0.0,#map1))) (all)
   do (#map2=ft(temp)) (all)

   {* Define macromolecule and solvent regions of the map *}
   {* compute cutoff *}
   Histogram
     mbins=999
     solcon=$solcon
     from=#map2                                  {* map2 is the smoothed mir map *}
   end
   evaluate ($cutoff=$result)

   do (map3=#map2) (all)
   do (#map2=1) (map3 < $cutoff)
   do (#map2=0) (map3 >= $cutoff)

   {*map2 =1 (< $cutoff)  represents the solvent region *}
   {*map2 =0 (>= $cutoff) represents the protein region *}
```

The initial map (default name `map`) is passed to the macro as object `#map1` and the mask (a map restricted to values of 1 and 0) returned to the main level of the *X-PLOR* script as `#map2` (default name `mask`). The size of the smoothing radius, `$Rs`, and the solvent content, `$solcon`, are defined at the beginning of the solvent-flattening script.

References

ABRAHAMS, J. P., LESLIE, A. G. W., LUTTER, R. & WALKER, J. E. (1994). *Nature (London)*, **370**, 621–628.

BAKER, D., BYSTROFF, C., FLETTERICK, R. J. & AGARD, D. A. (1993). *Acta Cryst.* D**49**, 429–439.

BLOW, D. M. & MATTHEWS, B. W. (1973). *Acta Cryst.* A**29**, 56–62.

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

BRÜNGER, A. T. (1992*a*). *Nature (London)*, **355**, 472–474.

BRÜNGER, A. T. (1992*b*). *X-PLOR Version* 3.1. *A System for X-ray Crystallography and NMR.*, Yale Univ. Press, New Haven, CT, USA.

BRÜNGER, A. T. (1993). *Acta Cryst.* D**49**, 24–36.

BRÜNGER, A. T. (1995). *Methods Enzymol.* In the press.

GIACOVAZZO, C., SILIQI, D. & RALPH, A. (1994). *Acta Cryst.* A**50**, 503–510.

GRIMES, J. & STUART, D. (1994). Proceedings of the Study Weekend organised by *CCP*4, pp. 67–76.

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

HSU, I.-N., DELBARE, L. T. J., JAMES, M. N. G. & HOFMANN, T. (1977). *Nature (London)*, **266**, 140–145.

JAMES, M. N. G. & SIELECKI, A. R. (1983). *J. Mol. Biol.* **163**, 299–361.

JIANG, J.-S. & BRÜNGER, A. T. (1994). *J. Mol. Biol.* **243**, 100–105.

JONES, T. A., & KJELDGÅARD, M. (1993). *O – The Manual, Version* 5.9, Uppsala Univ., Sweden.

LESLIE, A. G. W. (1988*a*). *A Reciprocal Space Algorithm For Calculating Molecular Envelopes Using The Algorithm of B. C. Wang. Improving Protein Phases.* Proceedings of the Study Weekend organised by *CCP*4, pp. 25–31.

LESLIE, A. G. W. (1988*b*). *Improving Protein Phases*, Proceedings of the Study Weekend organised by *CCP*4, pp. 13–24.

MATTHEWS, B. W. (1968). *J. Mol. Biol.* **33**, 491-497.

PFLUGRATH, J. W., WIEGAND, G., HUBER, R. & VÉRTESEY, L. (1986). *J. Mol. Biol.* **189**, 383-386.

PODJARNY, A. D., BHAT, T. N. & ZWICK, M. (1987). *Ann. Rev. Biophys. Biophys. Chem.* **16**, 351-373.
RAYMENT, I. (1983). *Acta Cryst.* A**39**, 102–116.
READ, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
RICE, D. W. (1981). *Acta Cryst.* A**37**, 491–500.
ROSSMANN, M. G. & BLOW, D. M. (1963). *Acta Cryst.* **16**, 39–45.
SCHEVITZ, R. W., PODJARNY, A. D., ZWICK, M., HUGHES, J. J. & SIGLER, P. B. (1981). *Acta Cryst.* A**37**, 669–677.
SIM, G. A. (1959). *Acta Cryst.* **12**, 813–815.
SRINIVASAN & PARTHASARATHY (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon Press.
STUART, D. & ARTYMIUK, P. (1985). *Acta Cryst.* A**40**, 713–716.
WANG, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
WOOLFSON, M. M. (1956). *Acta Cryst.* **9**, 804–810.
XIANG, S., CARTER, C. W., BRICOGNE, G. & GILMORE, C. J. (1993). *Acta Cryst.* D**49**, 193–212.
ZHANG, K. Y. J. (1993). *Acta Cryst.* D**49**, 213–222.
ZHANG, K. Y. J. & MAIN, P. (1990). *Acta Cryst.* A**46**, 41–46.