Acta Cryst. (1996). D52, 43–48

Phase Combination and Cross Validation in Iterated Density-Modification Calculations*

K. D. Cowtan[a] and P. Main[b]

[a]*Department of Chemistry and* [b]*Department of Physics, University of York, Heslington, York YO1 5DD, England*

(Received 22 May 1995; accepted 9 June 1995)

Abstract

A variety of density-modification techniques are now available for improving electron-density maps in accordance with known chemical information. This modification must, however, always be constrained by consistency with the experimental data. This is conventionally achieved by alternating cycles of map modification in real space with recombination with the experimental data in reciprocal space. The phase recombination is based upon the assumption that the density-modified map may be treated as a partial model of the structure which contains information independent of the experimentally derived phases. This assumption is shown to be incorrect, and an alternative procedure is investigated which as a side effect allows calculation of a free R factor.

1. Introduction

Density modification is conventionally applied iteratively, in a calculation of the following form.

The current phase estimates, are used to calculate an initial map by means of a Fourier transform. It is normal to weight the structure-factor amplitudes with the figures-of-merit attached to their phases.

Some technique is then applied to alter the map to 'improve' it on the basis of known chemical constraints or features of the particular structure. The altered map is back-transformed to produce a new set of structure-factor magnitudes and phases.

Some estimate is made of the error in the modified phase on the basis of the agreement between the observed and modified magnitudes. This error estimate is used to form a weighted combination of the phase information from the experimental phasing and the density modification. The phases are normally represented as probability distributions, which are combined by multiplication according to the assumption that the sources of phase information are independent.

The phase-probability distribution for the density-modified phase is conventionally generated under assumptions that were made for the combination of a partial atomic model with experimental data, *i.e.* that the calculated magnitudes and phases arise from a density map in which some atoms are present and correctly positioned, and the remainder are completely absent (Sim, 1959). Thus, the difference between the true structure factor and the calculated value must be the effective structure factor due to the missing density alone. If the phase of this quantity is random and the magnitude is drawn from a Wilson distribution (Wilson, 1949), then the difference between the magnitudes averaged over a set of reflections gives an estimate of the phase error.

This method and a modified form (the σ_A method, Read, 1986) are used almost universally in density-modification calculations. However in density-modification calculations the assumption that the experimental and density-modified map coefficients are independent sources of information is clearly invalid, since the experimental magnitudes, phases and weights are used in calculating the initial map to which the density modifications are applied. The error will be most serious in the case of modification methods which apply relatively small changes to the map, for example solvent flattening (Wang, 1985) and histogram matching (Zhang & Main 1990*a*,*b*). By contrast, iterative skeletonization (Baker, Bystroff, Fletterick & Agard, 1993) and high-order non-crystallographic symmetry averaging may modify the density dramatically, and so the problem will be less serious.

The process of density modification and phase combination is shown diagrammatically in Fig. 1.

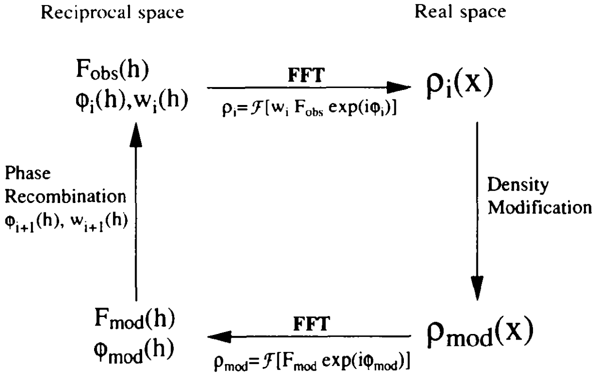

Fig. 1. The density-modification/phase recombination cycle. Symbols are defined in the text.

* A report of the meeting at which this paper was presented is published on pages 228–234 of this issue.

Acta Crystallographica Section D
ISSN 0907-4449 © 1996

2. Definitions

$\|F_{\text{obs}}\|$	Observed structure-factor magnitude
φ_{exp}	Experimental phase estimate, *e.g.* from MIR
w_{exp}	Estimated weight (figure-of-merit) for φ_{exp}
φ_i	*i*th cycle estimate of the phase of F_{obs}
w_i	Estimated weight (figure-of-merit) for φ_i
$\|F_{\text{mod}}\|$	'Structure-factor' magnitude calculated from modified map
φ_{mod}	Phase calculated from modified map
w_{mod}	Estimated weight (figure-of-merit) for φ_{mod}

3. What happens in extreme cases?

The problems with conventional phase combination algorithms can be shown by considering some extreme cases of density modification.

3.1. *'Perfect' density modification*

A perfect density-modification technique returns the experimental magnitudes (and phases) from any initial map. The perfect agreement between modified and observed magnitudes suggests that the phases are also perfect. Thus, weights for the modified phases w_{mod} should be 1, and the phase combination will return the modified phase.

Both the Sim and σ_A methods correctly give $w_{\text{mod}} = 1$.

3.2. *'Null' density modification*

A null density-modification technique returns the initial map, unmodified. The density-modification technique has no power for modifying the magnitudes towards their experimental values, thus the modified phases can also be assumed to be meaningless. Weights for the modified phases w_{mod} should be 0.

Both the Sim and σ_A methods return $w_{\text{mod}} \neq 0$. If iterated for many cycles, then $w_i \rightarrow 1$ as $i \rightarrow \infty$.

3.3. *'Random' density modification*

A random density-modification technique returns a random map from any initial map. Given prior knowledge of the structure individual reflections could be weighted according to the accuracy of their phases, but in a realistic calculation probably w_{mod} should be 0.

In this case the Sim method returns $w_{\text{mod}} \neq 0$, the σ_A method returns $w_{\text{mod}} = 0$.

The behaviour of current methods in the case of a 'null' or 'nearly null' modification method is particularly alarming. The density modification method is adding no information, and yet the figures of merit increase systematically due to the interdependence between the experimental and modified magnitudes.

4. Null density modification and Sim phase recombination

The case of 'null' density modification can be examined in more detail. The extreme case of a 'null' density-modification procedure is a procedure that produces no change in the initial map. Then, a 'nearly null' density-modification procedure is one that produces only a very small change in the map, and so must be iterated many times to obtain a significant phase improvement.

In the case of a null or nearly null density-modification technique, a map is calculated from the estimated phases. This modified map, which is almost identical to the initial map, is back-transformed. The resulting magnitudes are strongly correlated with the initial magnitudes, and thus the Sim weighting procedure produces strong weights for the modified data. Combination of the initial and final phases thus produces no change in the phases, while the figures-of-merit increase significantly.

Consider an acentric reflection **h**. In a phase-extension calculation the initial weight $w(\mathbf{h}) = 0$ for all **h** in the extension region. The initial map for cycle *i* is calculated with coefficients $F_i = w_i(\mathbf{h})|F_{\text{obs}}(\mathbf{h})|\exp[i\varphi_i(\mathbf{h})]$. Application of a 'nearly-null' density-modification procedure gives rise to a map whose coefficients are approximately the coefficients of the initial map, *i.e.*,

$$F_{\text{mod}} = w_i(\mathbf{h})|F_{\text{obs}}(\mathbf{h})|\exp[i\varphi_i(\mathbf{h})]. \qquad (1)$$

In calculating a weight for the modified phase, we need to estimate $\sum_Q$, the expected squared magnitude of the difference structure between the modified and true maps. In most variants of the Sim method this is acheived through an expression of the form,

$$\textstyle\sum_Q = \langle |F_{\text{obs}}|^2 - |F_{\text{mod}}|^2 \rangle. \qquad (2)$$

The weight for the modified phase is then estimated according to the equations,

$$w_{\text{mod}} = I_1(X)/I_0(X), \qquad (3)$$

where,

$$X = |F_{\text{mod}}||F_{\text{obs}}|/\textstyle\sum_Q. \qquad (4)$$

Thus, in the case where the figures of merit of most of the reflections (and thus the F_i and F_{mod}) are close to zero, we can make the approximation $\sum_Q = \langle |F_{\text{obs}}|^2 \rangle$,

$$\begin{aligned} X &= (2|F_{\text{mod}}||F_{\text{obs}}|)/\langle |F_{\text{obs}}|^2 \rangle \\ &= 2w_i(|F_{\text{obs}}|^2/\langle |F_{\text{obs}}|^2 \rangle). \end{aligned} \qquad (5)$$

$$\begin{aligned} w_{\text{mod}} &\simeq X/2 \text{ for small } X \\ &\simeq w_i(|F_{\text{obs}}|^2/\langle |F_{\text{obs}}|^2 \rangle). \end{aligned} \qquad (6)$$

In the extension region where there is no observed phase information the phase from the modified map and this weight will become the starting weight for the next cycle, w_{i+1},

$$w_{i+1} \simeq w_i(|F_{\text{obs}}|^2/\langle |F_{\text{obs}}|^2 \rangle). \qquad (7)$$

Thus, we can see that even in the case of a null density-modification procedure which adds no information, figures of merit in the phase-extension region will change in proportion to the square of the normalized structure-factor amplitude, thus those weights for which this is greater than one increase.

As the weights for some reflections increase, the corresponding values of $F_{\rm mod}$ will become significant and the estimated value of $\sum_Q$ will be less than $\langle|F_{\rm obs}|^2\rangle$, and so the weights will increase for more of the reflections. (3) prevents the values from exceeding 1.0.

This is clearly a problem. Ideally, if a null density-modification procedure is applied to a map, the phases and their weights should not change. The problem is essentially one of estimating how much independent information has been added in the density-modification process.

5. Approaches to the problem of over-consistency

Two possible approaches to the problem of over-consistency between observed and density-modified magnitudes are suggested.

(*a*) Reduce the number of cycles of density modification in which weakly phased reflections are included. Typically, density modification is started with only some subset of the data (for example, those reflections well phased from MIR data). Only these reflections are included in the phase recombination, with other reflections set to zero. As the calculation progresses, more reflections are introduced until all the data is included. The figures of merit of reflections which undergo fewer cycles of phase recombination will be correspondingly smaller.

This is an *ad hoc* method, based on some arbitrary assertion about which reflections will be accurately extrapolated and so should be included at an early stage and receive larger final figures of merit. However it has been the conventional approach in density-modification calculations (*e.g.* Leslie 1987; Zhang & Main 1990*a*,*b*).

(*b*) Make the modified map independent of the original map, (as was assumed when multiplying the phase-probability distributions in §1). This may be achieved through a reciprocal-space analogue of the omit map. Two algorithms have been examined.

5.1. *The 'reflection–omit' method*

The reflections are divided into ten groups, and density-modification calculations are performed excluding each set in turn as a free set. The reflections from each of the free sets are used combined to give a new complete data set which should be less dependent on the original magnitudes.

In the case of 'null' density modification the reflection–omit scheme will return all magnitudes equal to zero, thus $w_{\rm mod}=0$ for all reflections. However, each density-modification cycle must be repeated ten times to build up a full set of modified magnitudes. In the case of more complex density-modification schemes this may be an unacceptable cost.

5.2. *The 'free-Sim' method*

Instead of performing a full reflection–omit calculation, it is possible to gain some of the advantages for no extra computation by using a single omitted or 'free' set. This set of data is used in the estimation of $\sum_Q$ over each resolution shell [(2)]. This estimate is used in the calculation of weights for the rest of the reflections.

In the case of 'null' density modification the free-Sim method does not return $w_{\rm mod}=0$, however even if iterated for an infinite number of cycles the weights do not approach 1.

6. Problems with simple free-set calculations

A number of authors have used 'free-*R*' methods, developed for model refinement, in density-modification calculations, and the normal refinement procedure of excluding 5–10% of the data for the whole of the calculation has usually been adopted (*e.g.* Baker *et al.*, 1993). A free set of 5–10% is also required for the 'free-Sim' method, described above.

In molecular-replacement calculations, the model consists of the current coordinate estimates, temperature factors, form factors and so on. In density modification the 'model' consists of the current weighted map (*i.e.* the structure-factor magnitudes, phases and weights), and the density-modification constraints. Thus, the density-modification 'model' is already based on the structure-factor magnitudes.

The density-modification constraints can in general be written as relationships between structure factors in reciprocal space. If a set of magnitudes are omitted in the initial map, this becomes a part of the model, and as the calculation is iterated the phase estimates for the rest of the data will be modified into consistency with the zero free set magnitudes. The resulting bias in the map is worse than the bias resulting from a simple map calculation omitting 10% of the data.

The problem can be avoided in two ways.

(i) A different free set can be used for each cycle of density modification. This requires no additional computation, but a large noise component is introduced in the free *R* factor by the differences between free sets.

(ii) Each cycle of density modification can be repeated, one or more times with free sets to accumulate statistics for the free-*R* and free-Sim calculations. Finally, the cycle is repeated using all the reflections and phase combination is performed for all reflections.

Similar approaches are suggested in Brünger (1996) and Roberts & Brünger (1995). Note that neither of these methods would be considered allowable in the case of the refinement free R, since all the magnitudes are used at some stage in the calculation.

A simple test case demonstrates the phenomena. Phase improvement was performed on O^6-methylguanine–DNA methyltransferase (GMT, Moore, Gulbis, Dodson, Demple & Moody, 1994) using histogram matching and solvent flattening as density-modification constraints. After each cycle, the strength of extrapolation of the free set is measured by the r.m.s. magnitude of the free set. The calculation was performed using a single fixed free set (the conventional method), a different free set in each cycle (i above), and repeating the cycle with all the data (ii above). The results are plotted in Fig. 2 as a function of density-modification cycle.

We would expect that as the map improves the strength of extrapolation of the free set should increase. Using a single free-R set the strength of extrapolation increases for two cycles, and then starts falling as the working set reflections begin to reflect the absence of the free set. With either of the new schemes this problem does not occur and extrapolation increases towards a plateau. The noise caused by switching free-R sets (i above) is also apparent.

The weakening of extrapolation as the calculation continues is also reflected in the free-R factor (Fig. 3).

7. Some test results

Comparisons have been made using two structures. O^6-methylguanine–DNA methyltransferase (GMT, Moore *et al.*, 1994) has MIR phases to 3.0 Å and magnitudes to 2.4 Å. 5-Carboxymethyl-2-hydroxymuconate isomerase (CHMI, Wigley, Roper & Cooper, 1989) has weak MIR phases to 3.0 Å and magnitudes to 2.1 Å. The molecule packs with threefold non-crystallographic symmetry.

Three density-modification schemes were tested: solvent flattening (Wang, 1985) alone was applied to GMT; solvent flattening and histogram matching (Zhang & Main, 1990*a*,*b*), were applied to GMT; solvent flattening, histogram matching and non-crystallographic symmetry averaging were applied to CHMI.

In each of these cases the following phase combination methods are compared: conventional Sim phase combination using phase extension in resolution steps; 'free Sim' phase combination starting with all the data; and the reflection–omit procedure starting with all the data.

The quality of the modified maps, expressed as a map correlation coefficient with the map from the refined model, is plotted as a function of cycle number. Note that each cycle using the reflection–omit method takes ten times longer than a cycle using the other methods.

Note that in the case of solvent flattening alone (Fig. 4) or solvent flattening and histogram matching (Fig. 5) the conventional Sim method stops improving the map after a couple of cycles. Examination of the data shows that as predicted the figures of merit increase systematically towards 1.0 for all non-zero reflections, thus the weighted map is worse.

The free-Sim method produces a map which continues to improve, leading to a better final map.

The reflection–omit procedure produces a significantly better map than the other methods in five to ten cycles, after which the map begins to deteriorate. The free-R factor may be used to stop the calculation at this point. The map quality also oscillates from cycle to cycle, this is most likely an indication that the density-modification 'step' is too great and is overshooting the correct map. In the case of solvent flattening this will generally be true, and the use of a relaxation coefficient may lead to a stable solution.

The benefit of the new procedures can be seen in the case of solvent flattening by comparison of the phase errors and estimated weights (Table 1). Note that there is very little difference between the mean phase

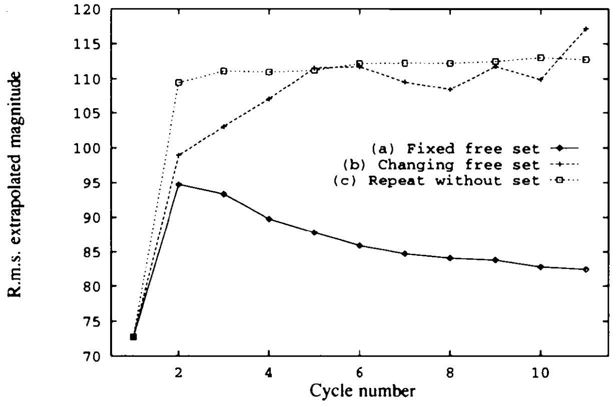

Fig. 2. Variation of the power of the extrapolation of free-set reflections as a function of density-modification cycle number. Case (*a*) fixed free set; case (*b*) changing free set (i in text); case (*c*) repeat without free set (ii in text).

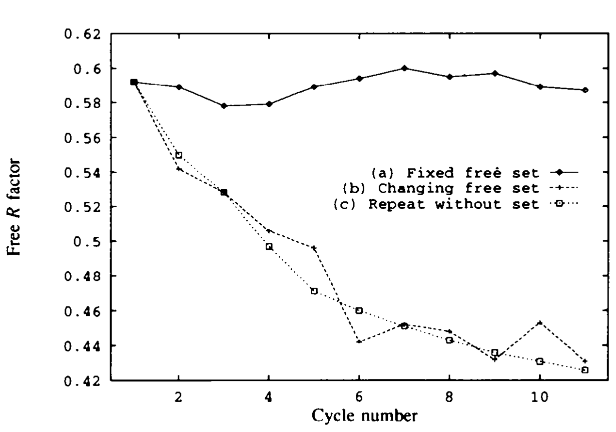

Fig. 3. Variation of free R factor as a function of density-modification cycle number. Case (*a*) fixed free set; case (*b*) changing free set (i in text); case (*c*) repeat without free set (ii in text).

error in each case, however the estimated weights for those phases are much smaller in the new methods. Calculating a line of regression between the cosines of the phase errors and the estimated weights reveals that the new methods, and in particular the reflection–omit method, lead to much more representative estimates for the weights.

In the case of non crystallographic symmetry averaging (Fig. 6), solvent flattening, and histogram matching, all the methods converge to almost identical final maps, and the difference in the final figures of merit between the methods is somewhat smaller. Convergence of the free-Sim method is slower than the conventional Sim calculation.

This suggests that histogram matching and solvent flattening provide comparatively weak information concerning the density, and care is needed in estimating the quality of the modified phases. By contrast, with threefold averaging the new information is quite strong and the phase weighting and combination process less critical.

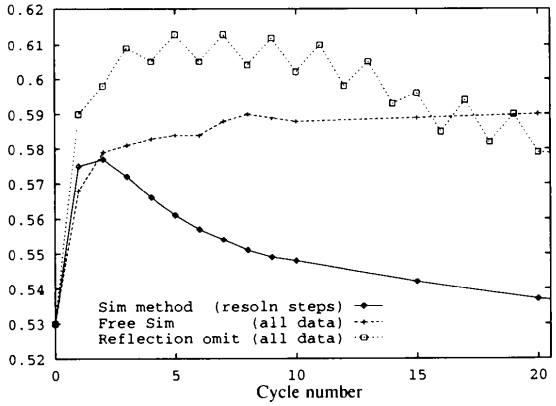

Fig. 4. Map quality as a function of density modification cycle for solvent flattening on GMT.

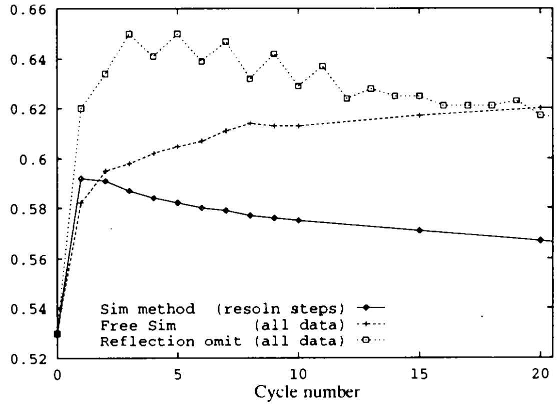

Fig. 5. Map quality as a function of density-modification cycle for solvent flattening and histogram matching on GMT.

Table 1. *Phase statistics after ten cycles of solvent flattening on GMT*

Weighted mean phase error is weighted with the figure-of-merit alone. $\Delta\varphi$ = phase error, w = figure-of-merit for each reflection.

Method	Unweighted mean $\Delta\varphi$	Weighted mean $\Delta\varphi$	Mean weight	Line of regression $\cos(\Delta\varphi) = mw + c$
Conv. Sim	72.9	68.7	0.75	$m = 0.39$ $c = -0.09$
Free Sim	72.9	66.2	0.57	$m = 0.48$ $c = -0.07$
Refl. Omit	73.9	58.8	0.33	$m = 0.71$ $c = -0.04$

The reflection–omit calculation does increase tenfold the computing time required for a density-modification calculation. However, a reflection–omit cycle may be introduced at any time in a density modification calculation, so a combined method involving several cycles of free-Sim combination followed by one cycle of reflection–omit may be a good compromise.

8. Concluding remarks

The application of conventional phase combination procedures as a part of density-modification calculations is based on some invalid assumptions which severely handicap the method in the case of 'weak' modification procedures, typically those which do not depend on any known structural information. Some problems are also to be expected in the case where the observed data used in phase recombination are very weak, for example when multi-crystal averaging with unphased crystal forms.

New phase combination methods are suggested to correct this problem. In the case of solvent flattening and histogram matching, these methods lead to significantly better results.

When there is strong information with which to improve the density, such as non-crystallographic symmetry, the phase-combination procedure becomes less critical, and the methods suggested here are not required.

KDC would like to thank the *CCP*4 project for funding this research.

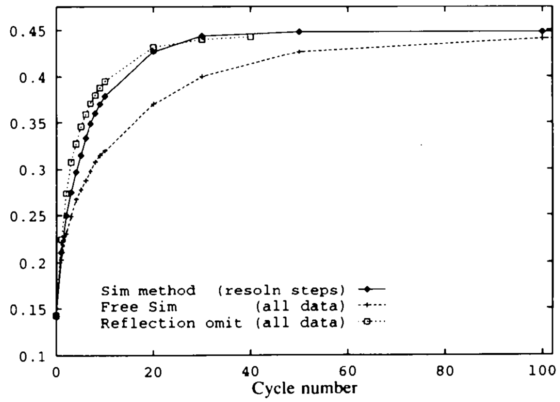

Fig. 6. Map quality as a function of density-modification cycle for NCS averaging, solvent flattening and histogram matching on CHMI.

References

Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–439.

Brünger, A. T. (1996). *Methods Enzymol.* In the press.

Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.

Moore, M. H., Gulbis, J. M., Dodson, E. J., Demple, B. & Moody, P. C. E. (1994). *EMBO J.* **13**, 1495–1501.

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

Roberts, A. L. U. & Brünger, A. T. (1995). *Acta Cryst.* D**51**, 990–1002.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

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

Wigley, D. B., Roper, D. I. & Cooper, R. A. (1989). *J. Mol. Biol.* **210**, 881–882.

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

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.