Acta Cryst. (1996). D**52**, 252–256

Direct-Space Methods in Phase Extension and Phase Refinement. IV. The Double-Histogram Method

L. S. Refaat, C. Tate and M. M. Woolfson

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

(*Received 5 May 1995; accepted 22 June 1995*)

Abstract

In the conventional histogram-matching technique for phase extension and refinement for proteins a simple one-to-one transformation is made in the protein region to modify calculated density so that it will have some target histogram in addition to solvent flattening. This work describes an investigation where the density modification takes into account not only the current calculated density at a grid point but also some characteristic of the environment of the grid point within some distance R. This characteristic can be one of the local maximum density, the local minimum density or the local variance of density. The grid points are divided into ten groups, each containing the same number of grid points, for ten different ranges of value of the local characteristic. The ten groups are modified to give different histograms, each corresponding to that obtained under the same circumstances from a structure similar to the one under investigation. This process is referred to as the double-histogram matching method. Other processes which have been investigated are the weighting of structure factors when calculating maps with estimated phases and also the use of a factor to dampen the change of density and so control the refinement process. Two protein structures were used in numerical trials, RNAp1 [Bezborodova, Ermekbaeva, Shlyapnikov, Polyakov & Bezborodov (1988). *Biokhimiya*, **53**, 965–973] and 2-Zn insulin [Baker, Blundell, Cutfield, Cutfield, Dodson, Dodson, Hodgkin, Hubbard, Isaacs, Reynolds, Sakabe, Sakabe & Vijayan (1988). *Philos. Trans. R. Soc. London Ser. B*, **319**, 456–469]. Comparison of the proposed procedures with the normal histogram-matching technique without structure-factor weighting or damping gives mean phase errors reduced by up to 10° with map correlation coefficients improved by as much as 0.14. Compared to the normal histogram used with weighting of structure factors and damping, the improvement due to the use of the double-histogram method is usually of order 4° in mean phase error and an increase of 0.06–0.08 in the map correlation coefficient. It is concluded that the most reliable results are found with the local-maximum condition and with R in the range 0.5–0.6 Å.

1. Introduction

The distribution of electron density has been found to be a sensitive and useful indicator of the errors in the phases used to produce the map. Thus its use, in the process known as histogram matching, has become a central feature in some methods of phase extension and refinement for proteins, for example in the widely used *SQUASH* procedure (Zhang & Main, 1990*a*,*b*) where, in conjunction with solvent flattening (Wang, 1985) and application of the Sayre equation (Sayre, 1972, 1974), it operates successfully at resolutions higher than about 4 Å or so. A quite different application of histogram matching has been made by Lunin (1993) who generated large numbers of very low resolution (8–10 Å) maps for proteins with random phases and recognized the more plausible maps by their histograms. It is interesting that histogram matching is capable of operating either at the coarse level of giving a very low resolution image of the structure or at the refined level of improving density to enable the fitting of a detailed model but, so far, it seems to have less to offer at intermediate resolutions.

2. The double histogram

In the normal histogram-modification procedure a one-to-one mapping is made in the protein region of old density ρ to new density ρ' such that ρ' will have the required histogram. This means that two grid points with the same value of ρ will also have the same value of ρ' so that the pattern of peaks and troughs in the modified map is similar to that in the original map. This coupling between the original and modified maps is broken when phases of the modified map are used with observed magnitudes and, under favourable conditions, the original pattern of peaks and troughs evolves towards the true pattern. In *SQUASH*, where different criteria for density modification are used simultaneously, this correlation between the original and modified maps is still present but is less strong.

In a correct map, points with the same density can have very different environments and we have examined various ways of modifying density to match expected histograms which depend on some characteristic of the environment. The characteristics of the local environment we have explored are, (i) local maximum density, $\rho_{1\max}$, (ii) local minimum density, $\rho_{1\min}$, and (ii) local variance, V_1. By local in this context we mean within some distance R of the grid point in question. Since the modified density now depends on two quantities, both the original density at the point and also some other local

characteristic we refer to this as the double-histogram (DH) method.

The idea behind the DH-matching method, illustrated with the local maximum density, is as follows.

For each of the M independent grid points in the protein region of the current map $\rho_{1\max}$ is found. This is done by a direct search of all the grid points within a distance R of the central one. For the $M/10$ grid points with the highest values of $\rho_{1\max}$ a histogram of the values of ρ is formed. This is repeated for the $M/10$ grid points with the next highest values of $\rho_{1\max}$, and so on, until there are ten histograms, H_j, $j = 1$ to 10, where $j = 1$ corresponds to grid points with the highest local maximum density and $j = 10$ corresponds to those with the lowest. The same process is carried out with density calculated for a known or synthetic structure similar to that under investigation to give the ten histograms ${}^tH_j$. The DH matching process consists of modifying densities to convert each H_j to ${}^tH_j$. In addition to modifying density in the protein region, solvent density is flattened in the usual way.

The local variance, $V_1 = \langle \rho^2 \rangle_1 - \langle \rho \rangle_1^2$, can be found by the use of Fourier transforms. From a map giving ρ, ρ^2 can easily be found and also its Fourier transform, G. Two kinds of average were used to find the variance; the first was within a sphere of radius R with uniform weight everywhere (ball function), the Fourier transform of which is Q, and the other a sphere of radius R where the weight varies linearly from a maximum at the centre to zero at the surface (tent function), with Fourier transform Q_t. Both Q and Q_t can be found analytically and, scaled so that integration of the function over the sphere gives unity, they are given by,

$$Q(s) = 3/(4\pi^2 s^2 R^2)\{[\sin(2\pi Rs)/2\pi Rs] - \cos(2\pi Rs)\}, \quad (1)$$

and

$$Q_t(s) = 3/(2\pi^3 s^3 R^3)\big(\{[1 - \cos(2\pi Rs)]/\pi Rs\} - \sin(2\pi Rs)\big). \quad (2)$$

The local variance is then found from $\mathrm{FT(QG)} - [\mathrm{FT(QF)}]^2$ or similarly with Q_t.

The procedure we followed is as follows.

(1) For a model structure resembling the one under investigation produce the ten individual histograms for the local characteristic (LC) under investigation, $\rho_{1\max}$, $\rho_{1\min}$, or V_1 within the protein region for an E-map calculated with normalized structure factors $E(\mathbf{h})$.

(2) Calculate an E-map, ρ_e, with the current phase estimates. Where these phase estimates have come from a map obtained in a previous cycle the Fourier coefficients have been given a weight,

$$W(h) = \tanh[|E(\mathbf{h})_c E(\mathbf{h})_o|/2], \quad (3)$$

where subscripts c and o refer to calculated and observed quantities, respectively.

(3) Define the protein and solvent regions of the map by the Wang (1985) procedure. This is only carried out for every alternate cycle of the refinement process.

(4) For the LC being investigated find the ten individual histograms, as described previously, for the points in the protein region.

(5) Modify the density of each point, taking into account which of the ten histograms it is associated with so that each of the ten individual histograms is matched by the modified density, ρ_m. The details of the basic histogram-matching process have been given by Zhang & Main (1990*a*).

(6) A damping procedure is used to reduce the change of density in each cycle. The revised modified density is given by,

$$\rho' = (1 - c)\rho_1 + c\rho_m, \quad (4)$$

where ρ_1 is the density map with the original, *e.g.* isomorphous, phase estimates. As c is reduced in value from unity to zero so the amount of damping is increased.

(7) In the solvent region ρ_m at each point is made equal to the average of ρ_e in the solvent region.

(8) Fourier transform the resultant map to obtain new phase estimates.

(9) Repeat from (2) until the average magnitude of phase shift (AMPS) from the previous cycle is less than 1°.

The quality of the phase sets obtained by this process were judged by three different criteria.

(i) The mean absolute phase error $\langle |\Delta\varphi| \rangle$.

(ii) The E-weighted absolute phase error $\langle |\Delta\varphi| \rangle_E$.

(iii) The map correlation coefficient (MCC). This was calculated from the individual phase errors using the formula given by Lunin & Woolfson (1993). In judging our results it should be noted that the MCC for a particular estimated phase set was always given in relation to an ideal map with calculated phases for all the available observed data. Thus, if phase estimates are available only for the 1.9 Å resolution data for 2Zn-insulin the MCC is given in relation to an ideal map at 1.5 Å resolution. Other authors have chosen to refer the MCC to ideal maps at the current resolution of their phase estimates but we prefer to make comparison with the ultimate target map.

We now describe some of the results we have obtained and give the conclusions we draw from them.

3. Numerical trials

Trials were carried out on two known protein structures. These were as follows.

(1) RNAp1 (Bezborodova, Ermekbaeva, Shlyapnikov, Polyakov & Bezborodov, 1988) with space group $P2_1$, $a = 32.01$, $b = 49.76$, $c = 30.67$ Å, $\beta = 115.83°$, $Z = 2$. The asymmetric unit contains 808 non-H atoms (including five S) in the protein plus 83 ordered water

Table 1. *The results of using different histogram-matching techniques of phase refinement for RNAp1*

The initial phase errors were $\langle|\Delta\varphi|\rangle = 65.2°$ and $\langle|\Delta\varphi|\rangle_E = 65.3°$. The E-maps used in the refinement process were unweighted. The normal histogram-matching process gave $\langle|\Delta\varphi|\rangle = 48.3°$, $\langle|\Delta\varphi|\rangle_E = 44.6°$ and MCC = 0.649.

DH		$R = 1.5$ Å			$R = 1.0$ Å			$R = 0.5$ Å	
method	$\langle\|\Delta\varphi\|\rangle$	$\langle\|\Delta\varphi\|\rangle_E$	MCC	$\langle\|\Delta\varphi\|\rangle$	$\langle\|\Delta\varphi\|\rangle_E$	MCC	$\langle\|\Delta\varphi\|\rangle$	$\langle\|\Delta\varphi\|\rangle_E$	MCC
$\rho_{1\max}$	47.3	43.5	0.663	46.9	43.1	0.668	46.5	42.7	0.673
$\rho_{1\min}$	48.1	44.3	0.652	49.4	45.7	0.634	50.8	47.1	0.618
V_1 (tent)	49.4	45.6	0.635	48.4	45.0	0.644	44.7	40.7	0.699
V_1 (ball)	50.4	46.7	0.623	48.9	45.0	0.645	46.5	42.6	0.675

Table 2. *The results of using different histogram-matching techniques of phase refinement for RNAp1*

The initial phase errors were $\langle|\Delta\varphi|\rangle = 65.2°$ and $\langle|\Delta\varphi|\rangle_E = 65.3°$. The E-maps used in the refinement process were weighted. The normal histogram-matching process gave $\langle|\Delta\varphi|\rangle = 42.5°$, $\langle|\Delta\varphi|\rangle_E = 38.4°$ and MCC = 0.744.

DH		$R = 1.5$ Å			$R = 1.0$ Å			$R = 0.5$ Å	
method	$\langle\|\Delta\varphi\|\rangle$	$\langle\|\Delta\varphi\|\rangle_E$	MCC	$\langle\|\Delta\varphi\|\rangle$	$\langle\|\Delta\varphi\|\rangle_E$	MCC	$\langle\|\Delta\varphi\|\rangle$	$\langle\|\Delta\varphi\|\rangle_E$	MCC
$\rho_{1\max}$	41.3	37.1	0.760	39.8	'35.5	0.775	39.9	35.6	0.774
$\rho_{1\min}$	40.7	36.6	0.765	42.5	38.4	0.745	44.4	40.4	0.717
V_1 (tent)	42.9	38.8	0.743	42.1	38.1	0.749	38.3	34.2	0.789
V_1 (ball)	43.4	39.2	0.737	42.6	38.4	0.746	39.6	35.3	0.779

molecules. There are 23 853 independent reflections to 1.17 Å resolution.

(2) 2Zn-insulin (Baker, Blundell, Cutfield, Cutfield, Dodson, Dodson, Hodgkin, Hubbard, Isaacs, Reynolds, Sakabe, Sakabe & Vijayan, 1988). Space group $R3$ with $a = 49.0$ Å and $\alpha_R = 114.8°$. The asymmetric unit contains 806 non-H atoms, excluding solvent but including two Zn atoms. There are 6450 reflections to 1.9 Å resolution, for which isomorphous-replacement phases estimates are available, and 13 289 reflections to 1.5 Å resolution.

For RNAp1 no isomorphous-replacement phase estimates were available so, to produce initial phases to refine φ_{mod}, we modified the calculated phases by applying,

$$\varphi_{\text{mod}} = \text{phase of}\{w \times \exp(i\varphi_R) + [1 - w]\exp(i\varphi_c)\}, \tag{5}$$

where φ_c is the calculated phase and φ_R is a random phase chosen from a uniform distribution between 0 and 2π. By varying w, phase sets can be produced varying continuously from calculated to random. In this case the initial phases we used had an unweighted mean phase error 65.24° (weighted 65.25°) and MCC = 0.337.

The first trial shown, in Table 1, is for RNAp1 without any damping, *i.e.* $c = 1$ in (4), and without using the weights $W(\mathbf{h})$, as given in (3). Various values of R have been tried for defining the LC. From these results, by comparison with those from the normal histogram-matching technique, it will be seen that for some value of R it is possible to obtain a better result with the DH for any of the three LC's but the best result is obtained with V_1, using the tent function, and with $R = 0.5$ Å. The errors, both weighted and unweighted, are almost 4° less and the MCC is higher by 0.05.

Table 3. *Normal histogram matching and solvent flattening with weighted Fourier syntheses and various values of c for 2Zn-insulin*

The 6450 isomorphous-replacement derived phases at 1.9 Å resolution had an unweighted and E-weighted mean phase errors of 59.0 and 56.4° and MCC = 0.358. The phase errors indicated in the table are for the complete data to the resolution indicated.

Damping factor c (Resolution, Å)	$\langle\|\Delta\varphi\|\rangle$	$\langle\|\Delta\varphi\|\rangle_E$	MCC
1.0			
(1.9)	49.5	43.4	
(1.5)	52.5	47.3	0.655
0.9			
(1.9)	44.9	39.9	
(1.5)	48.7	43.3	0.684
0.8			
(1.9)	45.8	41.7	
(1.8)	48.9	44.5	0.672

The same general conclusion may be drawn from Table 2 which shows a similar trial but with the use of weighted Fourier syntheses, where the weights are given by (3). The results are significantly better than those in Table 1 for the normal histogram procedure and for every type of DH. The extreme difference between using unweighted maps with a normal histogram and weighted maps with V_1 (tent function) and $R = 0.5$ Å amounts to a 10° reduction in phase error and an increase in MCC of 0.14. This improvement is very significant and shows the importance of the optimizations in histogram matching we have been exploring.

The 2Zn-insulin structure introduces the additional feature of phase extension. The 6450 reflections corresponding to 1.9 Å resolution, found by isomorphous replacement using a Hg derivative, started with an

Table 4. *Double-histogram matching results for 2Zn-insulin using the local maximum-density condition, solvent flattening and weighted Fourier syntheses for various values of the damping constant, c, and radius R*

The 6450 isomorphous replacement derived phases at 1.9 Å resolution had an unweighted and E-weighted mean phase errors of 59.0 and 56.4°, and MCC = 0.358. The phase errors indicated in the table are for the complete data to the resolution indicated.

	$R = 1.6$ Å			$R = 1.1$ Å			$R = 0.6$ Å		
c	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC
1.0									
(1.9 Å)	48.2	42.0		47.0	40.7		45.3	39.2	
(1.5 Å)	51.4	46.2	0.670	48.6	43.1	0.708	46.8	41.3	0.730
0.9									
(1.9 Å)	43.5	38.5		42.6	37.4		41.3	36.2	
(1.5 Å)	47.2	42.4	0.702	45.8	40.8	0.723	43.8	38.9	0.745
0.8									
(1.9 Å)	44.6	40.3		44.2	39.8		42.8	38.4	
(1.5 Å)	47.7	43.1	0.690	46.5	41.8	0.708	44.7	40.0	0.725

Table 5. *Double-histogram matching results for 2Zn-insulin using the local variance (tent function) condition, solvent flattening and weighted Fourier syntheses for various values of the damping constant, c, and radius R*

The 6450 isomorphous replacement derived phases at 1.9 Å resolution had an unweighted and E-weighted mean phase errors of 59.0 and 56.4°, and MCC= 0.358. The phase errors indicated in the table are for the complete data to the resolution indicated. With $R = 0.5$ Å the outcome was random phases for all situations

	$R = 2.0$ Å			$R = 1.5$ Å			$R = 1.0$ Å		
c	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC
1.0									
(1.9 Å)	45.4	39.1		46.7	40.7		51.0	44.7	
(1.5 Å)	47.6	42.1	0.743	49.5	44.2	0.694	54.9	49.8	0.626
0.9									
(1.9 Å)	41.5	36.7		42.6	37.8		46.7	41.4	
(1.5 Å)	44.5	39.6	0.762	45.7	40.9	0.721	52.3	47.9	0.635
0.8									
(1.9 Å)	43.1	38.9		43.6	39.5		46.7	42.3	
(1.5 Å)	45.1	40.5	0.745	46.1	41.7	0.708	51.9	47.7	0.634

unweighted mean phase error of 59.0°, a weighted mean phase error of 56.4° and an MCC = 0.358. The MIR phases were refined by the histogram-matching process under test until the AMPS was less than 3.5°. Then new reflections were added to the system, increasing the resolution 0.1 Å each time, and refined until the AMPS was less than 2°. Finally, once all reflections to 1.5 Å resolution were included, the refinement was continued until the AMPS was less than 1°.

Table 3 shows the result of using the normal histogram-matching process. The weights given by (3) were used in calculating the E-maps and various values of c were also tried. Our tests indicated that c had to be fairly close to unity so we show the results only for $c = 0.8$, 0.9 and 1.0. The table shows the final mean phase errors, weighted and unweighted, for both the 1.9 Å resolution subset of reflections and for the complete set of 13 289 reflections to 1.5 Å resolution and the MCC for only the complete set. It is clear that the refinement process has been quite successful and that the final map would be readily interpretable.

Table 4 shows the results of the DH method using $\rho_{1\max}$ as the LC with various values of c and R. It is evident that, for any particular value of c, all the values of R chosen gave better final phases than did the normal histogram-matching method. The best result, for $c = 0.9$ and $R = 0.6$ Å, was about 4° better in terms of mean phase error, either weighted or unweighted, and 0.06 better in MCC than that obtained from the normal histogram-matching procedure.

The best results in terms of MCC were obtained with the tent-function version of V_1 as the LC and these are seen in Table 5. However, in contrast to what is seen in Table 2 it is the larger values of R which give the better results in this case. While we are not completely certain of the cause of the difference we feel that it is likely to be due to the presence of the heavy atoms in 2Zn-insulin. The regions around the Zn atoms are going to be the regions of largest local variance for 2Zn-insulin and no such systematic regions exist for RNAp1. This is something that must be more thoroughly explored since results depend quite critically on the selected value of R. The best result from Table 5, which is for $c = 0.9$, is again, when compared with the normal histogram-matching procedure, about 4° better in terms of mean phase error, either weighted or unweighted, but 0.08 better in MCC in this case.

For completeness we give the results of using $\rho_{1\min}$, as the LC but even the best results in this case are slightly worse than using the normal histogram method.

Table 6. *Double-histogram matching results for 2Zn-insulin using the local minimum condition, solvent flattening and weighted Fourier syntheses for various values of the damping constant, c, and radius R*

The 6450 isomorphous-replacement derived phases at 1.9 Å resolution had an unweighted and E-weighted mean phase errors of 59.0 and 56.4°, and MCC = 0.358. The phase errors indicated in the table are for the complete data to the resolution indicated.

		$R = 1.6$ Å			$R = 1.1$ Å			$R = 0.6$ Å	
c	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC	$\langle\lvert\Delta\varphi\rvert\rangle$	$\langle\lvert\Delta\varphi\rvert\rangle_E$	MCC
1.0									
(1.9 Å)	51.6	45.3		55.7	49.9		66.6	61.8	
(1.5 Å)	54.2	48.9	0.638	58.8	54.1	0.566	74.9	72.1	0.303
0.9									
(1.9 Å)	45.3	40.2		48.9	43.9		58.8	53.9	
(1.5 Å)	49.2	44.4	0.681	54.6	50.2	0.609	70.7	68.0	0.351
0.8									
(1.9 Å)	46.7	42.4		49.1	44.8		56.6	52.0	
(1.5 Å)	49.7	45.1	0.667	54.6	50.5	0.598	68.6	66.0	0.379

4. Concluding remarks

A firm conclusion we have reached is that judicious use of the DH-matching method can give appreciably better results than using the normal histogram-matching procedure. The choice of the damping factor c is consistently indicated as about 0.9 but the best value for the averaging radius R seems to be structure dependent. Good results are obtained with either $\rho_{1\max}$ or the tent-function version of V_1 as the LC. However, the most reliable choice of parameters seems to be to use $\rho_{1\max}$ with $c = 0.9$ and a value of R in the range 0.5–0.6, which gives good results for both RNAp 1 and 2Zn-insulin.

We have illustrated the effectiveness of the DH-matching procedure used on its own but it is usually used in conjunction with other procedures for phase extension and refinement. It is being incorporated into a package *PERP* (phase extension and refinement program) which will also contain, solvent flattening, a novel Sayre-equation refinement algorithm (Refaat, Tate & Woolfson, 1996) and low-density elimination (Shiono & Woolfson, 1991; Refaat & Woolfson, 1993). Also to be included is a histogram-matching process based on a principle completely different from any used hitherto (Gu, Yao & Woolfson, 1996), which it is hoped will reinforce the combination of methods described above.

We are pleased to thank the Science and Engineering Research Council for their generous support of this project and for other related activity.

References

Baker, E. N., Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. M. C., Hubbard, R. E., Isaacs, N. W., Reynolds, C. D., Sakabe, K., Sakabe, N. & Vijayan, N. M. (1988). *Philos. Trans. R. Soc. London Ser. B*, **319**, 456–469.

Bezborodova, S. I., Ermekbaeva, L. A., Shlyapnikov, S. V., Polyakov, K. M. & Bezborodov, A. M. (1988). *Biokhimiya*, **53**, 965–973.

Gu, Y.-X, Yao Jia-X. & Woolfson, M. A. (1996). In preparation.

Lunin, V. Yu. (1993). *Acta Cryst.* D**49**, 90–99.

Lunin, V. Yu. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1996). In the press.

Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.

Sayre, D. (1972). *Acta Cryst.* **28**, 210–212.

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

Shiono, M. & Woolfson, M. M. (1992). *Acta Cryst.* A**48**, 451–456.

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

Zhang, K. M. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 41–46.

Zhang, K. M. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 377–381.