Acta Cryst. (1990). A**46**, 41-46

Histogram Matching as a New Density Modification Technique for Phase Refinement and Extension of Protein Molecules

BY KAM YONG JIAN ZHANG AND PETER MAIN

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

(*Received* 14 *April* 1988; *accepted* 11 *August* 1989)

Abstract

A new density modification technique - histogram matching - is being developed with encouraging results. Its application to the known structure of pig 2Zn insulin refines the 6500 1·9 Å MIR phases from a mean error of 60° to one of 46°. With these refined phases as a starting point for phase extension to 1·5 Å, the mean error for the final 13 000 phases is 46°. The original 1·9 Å phases continue to refine during the phase extension to a final mean error of 40°. A comparison is made with similar calculations already published.

Introduction

Phase refinement and extension play an important role in the determination of macromolecular structures. The initial phases, most commonly available from MIR, may not be sufficiently accurate to give a readily interpretable map or may only be available at low resolution despite the existence of high-resolution data for the native crystal. It is therefore of obvious importance to be able to refine and extend the MIR phases to produce an interpretable map at the full resolution of the native data. This may be done either in real space (density modification) or reciprocal space (direct methods) or a combination of the two. For a recent review of density modification methods see Podjarny, Bhat & Zwick (1987).

We describe in this paper a density modification technique, known as histogram matching, that we have applied to the known structure of *R*3 2Zn insulin (Baker *et al.*, 1989) with encouraging results. It achieves both phase refinement and extension at modest computing cost and the results are compared with those obtained from the following methods:

(i) Agarwal & Isaacs (1977) described a dummy-atom refinement method which they applied to the structure of insulin. This consists of fitting dummy atoms to the approximate map and then refining the positional and thermal parameters by least squares to minimize the discrepancy between observed and calculated magnitudes.

(ii) Direct phase extension and refinement have been achieved by Sayre (1972) using Sayre's equation, which he applied successfully to the structure of rubredoxin (Sayre, 1974). The same method was applied to insulin and reported by Agarwal & Isaacs (1977).

(iii) Another reciprocal-space method that has been successful is the maximum determinant rule of Tsoucaris (1970). The application of this to the structure of insulin is reported by de Rango, Maugen, Tsoucaris, Dodson, Dodson & Taylor (1985).

(iv) The method of solvent flattening was described by Wang (1985) as a means of solving the phase ambiguity of single-isomorphous-replacement or single-anomalous-scattering data. It is now widely used in the refinement of MIR phases and we report here its application to 2Zn insulin.

Histogram matching is a standard technique in image processing [see, for example, Castleman (1979), chapter 6] but is applied here for the first time to X-ray crystallography. It may be regarded as a generalization of solvent flattening and is also related to the 'phase correction' technique of Hoppe & Gassmann (1968). For any discrete image, a histogram of density values can be obtained. In many cases this can be compared with the histogram expected of a good image and used as a measure of quality. Furthermore, the test image may be improved by adjusting its density values in a systematic way to make its histogram match the correct histogram exactly. In order to apply this to X-ray crystallography, we need to predict the density histogram of the electron density map of an unknown structure and establish that this differs from the histogram of an approximate map. We then modify the density values of the approximate map to match the correct histogram and calculate new phases from the modified map. This forms one cycle of an iterative process of map improvement.

Method

(i) *Histogram matching*

The density histogram of an electron density map is the probability distribution of electron density values at the grid points at which the map is evaluated. It provides a global description of the appearance of the map and all spatial information is discarded. The process of histogram matching transforms the present

0108-7673/90/010041-06$03.00

electron density distribution into the distribution expected of a good map. This is done as follows:

(*a*) Compute the histogram of the map to be modified and acquire the expected histogram at the same resolution. The latter may be taken from the map of a similar structure or calculated from a formula (see later).

(*b*) Divide the two histograms into equal areas, as shown in Fig. 1. This gives corresponding density values ρ_i and ρ'_i, $i = 1, \ldots, n$ in the two histograms. A value of n of about 100 is quite satisfactory.

(*c*) From these density values, calculate scale factors a_i and shifts b_i which will map ρ onto ρ' within the interval ρ_i to ρ'_i as

$$\rho' = a_i\rho + b_i, \qquad (1)$$

i.e.

$$a_i = \frac{\rho'_{i+1} - \rho'_i}{\rho_{i+1} - \rho_i} \quad \text{and} \quad b_i = \frac{\rho'_i\rho_{i+1} - \rho'_{i+1}\rho_i}{\rho_{i+1} - \rho_i}.$$

It can be seen that $b_i > 0$ shifts the histogram to the right, while $b_i < 0$ shifts it left. Also, $a_i > 1$ broadens the histogram while reducing its height to keep the area constant.

(*d*) Apply the operation in (1) to the map, using the appropriate values of a_i and b_i for each range of ρ. This produces a new map which has the same electron density distribution as the expected one. Note that this operation also applies a maximum and a minimum value to the electron density, imposes the correct mean and variance and defines the entropy of the new map.

(ii) *Phase refinement and extension*

We have combined histogram matching with Wang's solvent flattening technique to produce the following iterative procedure of map improvement. Starting from an approximate map calculated from MIR phases:

(*a*) Determine the molecular envelope as described by Wang.

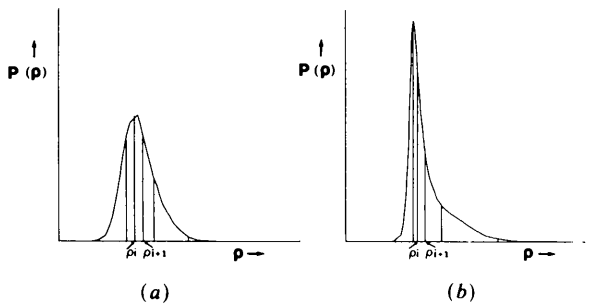


Fig. 1. (*a*) Electron density histogram from an approximate map. The area is divided into equal smaller areas with boundaries at ρ_i, $i = 1, \ldots, n$. (*b*) Expected histogram divided into equal areas as in (*a*) with boundaries at ρ'_i, $i = 1, \ldots, n$. This gives corresponding density values ρ_i and ρ'_i, $i = 1, \ldots, n$ in the two maps.

(*b*) Set the density within the solvent region to an appropriate constant.

(*c*) Obtain the expected histogram at the desired resolution. This will be a higher resolution than the present map if phase extension is required.

(*d*) Modify the map within the molecular envelope to match the expected histogram.

(*e*) Transform the modified map to obtain structure factors at the new resolution and calculate their Sim weights.

(*f*) The new phases are combined with the original MIR phases, taking their weights into account. The extended phases and weights are accepted at their calculated values.

(*g*) Calculate a new map and repeat from (*a*) until the process has converged.

Experimental results

(i) *Density histograms of actual protein maps*

The density histograms of a small number of protein and protein-like structures were examined and they were all found to behave in the same way. They were independent of the grid size on which the map was calculated, provided this was fine enough to give a good representation of the underlying continuous density. They were also independent of the structure itself, showing that a histogram for a known structure could be used for an unknown structure. However, the histograms changed with resolution and were also dependent on the overall temperature factor.

Fig. 2 shows the histograms from two different proteins at a range of resolutions. The densities were taken only from within the molecular envelope, the volume of which was calculated to give an average of 10 Å^3 for the protein atoms (including hydrogen). This ensured the same ratio of atomic volume to background in all cases. It can be seen from Fig. 2

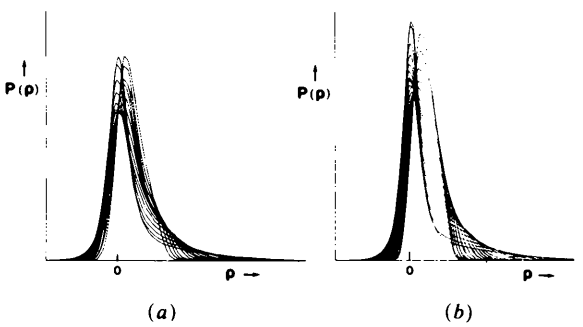


Fig. 2. (*a*) Electron density histograms of pig 2Zn insulin at resolutions ranging from 1·5 to 4·1 Å obtained from maps given by refined atomic coordinates. (*b*) As in (*a*), but the histograms are for haemoglobin (Derewenda, Dodson, Dodson & Bizozowski, 1981). The high-resolution maps give rise to the high peak near $\rho(x) = 0$. The corresponding peak for low-resolution maps is broader with a maximum at a higher value of $\rho(x)$.

that, as the resolution decreases, the peak of the histogram lowers to a minimum at about 3·0 Å and then rises again. As the peak rises, its maximum moves towards higher density and all the time the peak becomes broader. This large peak is due to the low density in the space between the atoms. The space becomes smaller with decreasing resolution and eventually disappears. Low density then becomes even more scarce with increasing atomic overlap at the lower resolution, pushing the mode of the distribution to higher values. The atoms in the map give rise to the long low tail of the histogram stretching out to high density values. This tail contracts as the resolution decreases and the atoms become more spread out.

It should be noted that there is nothing in this explanation of histogram behaviour that depends upon the details of the structure. The shape of the histogram depends only upon the fact that atoms are present and are at certain characteristic distances apart. This will be true for all polypeptide structures. Because of this, the histograms used in the present work were all taken from maps calculated from the refined atomic coordinates of pig 2Zn insulin. These will differ little from histograms taken from any other similar source.

In order to use the histogram information, electron density maps calculated from approximate phases should have histograms which differ from those expected. This is seen in Fig. 3 which compares the histograms of the 1·9 Å maps of pig 2Zn insulin calculated from refined atomic coordinates and from the isomorphous phases. There is a considerable difference between them, indicating the possibility of map improvement by the process of histogram matching.

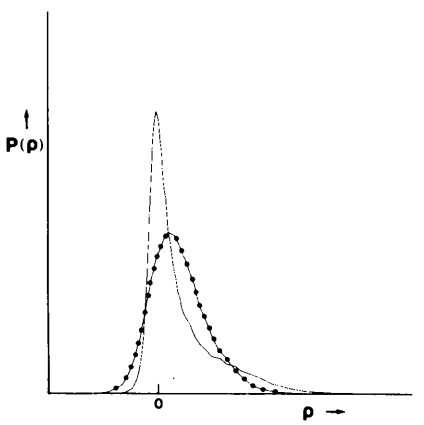


Fig. 3. Electron density histograms of pig 2Zn insulin at 1·9 Å resolution. –·–·– Map obtained from MIR phases. ——— Map obtained from refined atomic coordinates.

(ii) *Phase refinement and extension*

The starting point for tests of map improvement was the 1·9 Å map of pig 2Zn insulin calculated from phases obtained by multiple isomorphous replacement. It is already of good quality, for the structure was obtained from an earlier MIR map at 2·8 Å resolution. However, the same data have been used in tests of other methods, as described in the introduction, allowing us to compare results. The space group is $R3$ and the asymmetric unit contains 806 non-hydrogen atoms belonging to the protein. There are also two Zn atoms in the cell.

The initial MIR phases were refined using the procedure outlined in the previous section and convergence was reached after five iterations. The phases were then extended from 1·9 to 1·5 Å in four stages, increasing the resolution by 0·1 Å at each stage. Each time, convergence was reached after five iterations. The results are shown in Table 1. The two measures of quality used here are the mean phase error over all reflexions within the indicated resolution range and the correlation coefficient between the map and that obtained from the known atomic coordinates at the same resolution. The correlation coefficient is calculated from

$$\frac{\overline{\rho(x)\rho'(x)} - \overline{\rho(x)}\,\overline{\rho'(x)}}{[\overline{\rho^2(x)} - \overline{\rho(x)}^2]^{1/2}[\overline{\rho'^2(x)} - \overline{\rho'(x)}^2]^{1/2}}$$

where $\rho(x)$ and $\rho'(x)$ represent the density values in the two maps and $\overline{\rho(x)}$ is a mean value calculated over the whole map. Weights are applied in the calculation of the maps, but no weights are used in the calculation of mean phase error.

There is a clear improvement in the quality of the map as indicated by the correlation coefficient. In terms of the phases, the MIR phases refine from an initial mean error of 60 to 47°, then continue to improve to 42° mean error during the phase extension. The newly extended phases are determined to a better accuracy than the original MIR phases.

Further tests were carried out to see if additional improvement could be obtained by using sharpened F's, or even normalized structure amplitudes, instead of the F_{obs} used so far. The most accurate phases were obtained using F's from which the overall temperature factor had been removed, *i.e.* F's corresponding to stationary atoms. The results of this phase refinement and extension are set out in Table 2. Compared with the previous results, they show a mean improvement of 2° for the original phases and 3° for the extended phases.

It may be seen that the 1·9 Å phases continue to improve as more magnitudes are included in the map. This gives rise to the possibility of further improvement by repeating the phase extension using the more highly refined 1·9 Å phases as a starting point. Table 3 shows the results of this. An additional improvement

Table 1. *Result of map improvement by histogram matching for* 2Zn *insulin*

Resolution (Å)	Number of reflexions	Mean phase error (°) 1·9 Å phases	Extended phases	All phases	Correlation coefficient
1·9	6537	59·9	—	59·9	0·668
1·9	6537	47·0	—	47·0	0·778
1·5	13 287	42·3	54·8	48·4	0·803

Table 2. *Result of map improvement for* 2Zn *insulin using sharpened F's*

Resolution (Å)	Number of reflexions	Mean phase error (°) 1·9 Å phases	Extended phases	All phases	Correlation coefficient
1·9	6537	59·9	—	59·9	0·589
1·9	6537	46·3	—	46·3	0·728
1·8	7657	44·2	57·0	46·1	0·746
1·7	9130	42·6	55·4	46·2	0·754
1·6	10 946	41·3	52·4	44·9	0·757
1·5	13 287	40·1	51·8	45·9	0·756

Table 3. *Repeat of phase extension starting from refined* 1·9 Å *phases*

Resolution (Å)	Number of reflexions	Mean phase error (°) 1·9 Å phases	Extended phases	All phases	Correlation coefficient
1·9	6537	40·1	—	40·1	0·771
1·5	13 287	38·8	49·9	44·4	0·773

Table 4. *Comparison of mean phase errors* (°) *for* 2Zn *insulin obtained by different methods*

Resolution (Å)	MIR	Dummy-atom refinement*	Sayre's equation	Maximum determinant	Solvent flattening	Histogram matching
1·9	60	65	52	49	45	39
1·5	—	70	55†	52	52	44

* The dummy-atom refinement was started from 3·0 Å MIR phases.
† Only 10 000 phases were included in this mean error.

Table 5. *Mean phase error* (°) *of the strongest reflexions*

The MIR phases are at 1·9 Å resolution, the remainder are all at 1·5 Å.

Number of strongest reflexions	MIR phases	Dummy-atom refinement	Solvent flattening	Histogram matching
250	31	27	18	16
500	33	32	20	17
1000	37	39	23	20
2000	46	47	30	24

is indeed achieved, though probably not worth the doubling of computer time it entails.

Discussion

It was mentioned previously that the density histogram depends upon the overall temperature factor. However, we have seen that the best phases are obtained after removing the effects of temperature from the F's. If magnitudes sharpened in this way are always used, it becomes unnecessary to change the histogram according to the temperature factor of different structures, thus simplifying the use of the method.

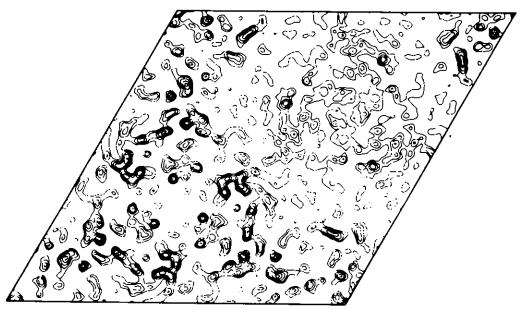
It is satisfying to note that the 1·9 Å phases continue to improve as magnitudes at higher resolution are added to the map. It is also satisfying to find that the extended phases are more accurate than the original MIR phases. A comparison of phase errors with those obtained by the methods mentioned in the *Introduction* is shown in Table 4. The solvent flattening was carried out by the present authors using a volume of 30% of the cell for the solvent. All other results were obtained by the authors referenced. Table 5 gives a closer comparison with the dummy-atom refinement method and Table 6 details the comparison with the maximum determinant results.

It is clear that the histogram matching method (which incorporates solvent flattening) produces the best results. In addition, it requires very much less computing time than all the other methods considered except solvent flattening. This is because the most substantial part of the computation is the calculation of Fourier transforms.

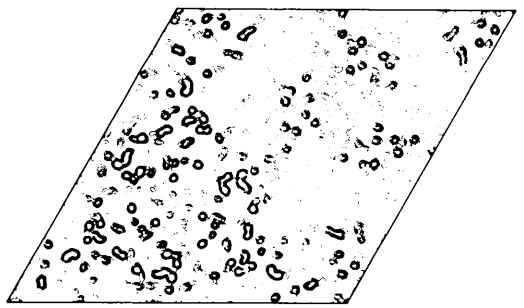
Table 6. *Mean phase error* (°) *as a function of E value*

Resolution (Å)	Number of reflexions	$\|E\|$	MIR phases	Maximum determinant	Solvent flattening	Histogram matching
1·9	604	>1·5	57	27	29	20
1·5	1147	>1·5		28	31	22
1·9	2408	>1·0	54	33	32	25
1·5	5020	>1·0		38	40	31
1·9	6522	>0·1	60	49	45	39
1·5	13 281	>0·1		52	52	44

We have demonstrated that the density histogram of an electron density map contains information which can be exploited in a process of map improvement at high resolution. When combined with solvent flattening, we have a method which restricts electron density values over the whole cell instead of just the solvent region or the molecule. Tests of the method at lower resolution were not as satisfactory as those already described. The initial isomorphous phases refined very well but, upon phase extension, the errors in the new phases rapidly became too large. Work is now in progress to improve on this. As was pointed out previously, the density histogram discards all positional information. Although the histogram is unique for any particular map, vastly different maps can have identical histograms. This makes histogram matching inherently less powerful than solvent flattening since, in the latter method, positional information is always available if the molecular envelope is known.



(*a*)



(*b*)

(*c*)

Fig. 4. (*a*) Section of insulin map calculated from 1·9 Å isomorphous phases. (*b*) Same section as (*a*), but calculated from 1·5 Å phases obtained from histogram matching. (*c*) Same section as (*a*), but obtained at 1·5 Å resolution from refined atomic coordinates.

Histogram matching, as applied here, suffers from the same defects as solvent flattening in that molecular density outside the envelope is strongly suppressed. To combat this problem, we are experimenting with new techniques of determining the envelope. In addition, false density inside the envelope tends to remain. It was observed in our tests on insulin that much false density was suppressed and new correct density appeared, so that a genuine improvement of the map was achieved. This may be judged from Fig. 4 which shows the same section from each of three maps – the original 1·9 Å isomorphous map, the 1·5 Å map obtained from histogram matching and the 1·5 Å map given by refined atomic coordinates. The molecular boundary used in the calculations is superimposed on the latter map.

We wish to thank Professor G. Dodson for kindly supplying the 2Zn insulin data and atomic coordinates. We are also indebted to Mrs E. Dodson for the use of computer programs and helpful discussions. One of us (KYJZ) is very grateful to the Rigaku Corporation of Japan for a research studentship and also to the Dodsons for the use of their laboratory facilities.

References

AGARWAL, R. C. & ISAACS, N. W. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 2835-2839.

BAKER, E. N., BLUNDELL, T. L., CUTFIELD, J. F., CUTFIELD, S. M., DODSON, E. J., DODSON, G. G., HODGKIN, D. C., HUBBARD, R. E., ISAACS, N. W., REYNOLDS, C. D., SAKABE, N. & VIJAYAN, M. (1989). *Philos. Trans. R. Soc. London.* In the press.

CASTLEMAN, K. R. (1979). *Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall.
DEREWENDA, Z. S., DODSON, E. J., DODSON, G. G. & BIZOZOWSKI, A. M. (1981). *Acta Cryst.* A**37**, 407-413.
HOPPE, W. & GASSMANN, J. (1968). *Acta Cryst.* B**24**, 97-107.
PODJARNY, A. D., BHAT, T. N. & ZWICK, M. (1987). *Annu. Rev. Biophys. Chem.* **16**, 351-373.
RANGO, C. DE, MAUGEN, Y., TSOUCARIS, G., DODSON, E. J., DODSON, G. G. & TAYLOR, D. J. (1985). *Acta Cryst.* A**41**, 3-17.
SAYRE, D. (1972). *Acta Cryst.* A**28**, 210-212.
SAYRE, D. (1974). *Acta Cryst.* A**30**, 180-184.
TSOUCARIS, G. (1970). *Acta Cryst.* A**26**, 492-499.
WANG, B. C. (1985). *Methods Enzymol.* **115**, 90-112.

Acta Cryst. (1990). A**46**, 46-57

Extension of Molecular Replacement: a New Search Strategy based on Patterson Correlation Refinement

BY AXEL T. BRÜNGER

The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, *USA*

(*Received* 28 *April* 1989; *accepted* 18 *August* 1989)

Abstract

A new search strategy is presented to obtain initial phases for single-crystal diffraction data by molecular replacement. It consists of carrying out 'Patterson refinements' of a large number of the highest peaks of a rotation function. The target function for Patterson refinement is proportional to the negative correlation coefficient between the squared amplitudes of the observed and the calculated normalized structure factors. If the root-mean-square difference between the search model and the crystal structure is within the radius of convergence of the minimization procedure employed, the correct orientation can be identified by having the lowest value of the target function after refinement. Similar to conventional crystallographic *R*-factor refinement, the target function for Patterson refinement may be combined with an empirical energy function describing geometric and non-bonded interactions. Patterson refinement of individual atomic coordinates or of rigid-group parameters may be carried out. Search models of crambin and of myoglobin with 1·6-2·0 Å backbone atomic r.m.s. differences from the target crystal structures show that the Patterson refinement strategy can solve crystal structures that cannot be solved by conventional molecular replacement or even by full six-dimensional searches.

Abbreviations

CPU, central processing unit; FFT, fast Fourier transformation; MR, molecular replacement; PC, standard linear correlation coefficient between $|E_{obs}|^2$ and $|E_{model}|^2$; RF, rotation function; r.m.s., root-mean-square; SA refinement, crystallographic refinement by simulated annealing with molecular dynamics; S/N, signal to noise; TF, translation function.

Introduction

In macromolecular crystallography, the initial determination of phases by 'molecular replacement' (MR) (Rossmann & Blow, 1962; Huber, 1965; Rossmann, 1972; Lattman, 1985) is often attempted if the structure of a similar or homologous macromolecule is known ('search model'). MR involves the placement (*i.e.* rotation and translation) of the search model in the unit cell of the target crystal in order to obtain the best agreement between calculated and observed diffraction data. The optimally placed search model is used to obtain initial phases for structure building and refinement. This approach may or may not succeed; many successful cases reported involve search models with a backbone atomic root-mean-square (r.m.s.) difference of less than 1 Å from the target structure (*e.g.* Wang, Bode & Huber, 1985; Schirmer, Huber, Schneider, Bode, Miller & Hackert, 1986).

Recent progress in obtaining approximate three-dimensional models of macromolecules from other information suggests an increased use of MR to solve crystal structures. For instance, the data base of known protein sequences and protein structures is growing rapidly. Techniques for aligning sequences such as consensus templates [see Taylor (1988) for a review] have been developed in order to recognize very distantly related proteins or protein domains and to carry out model building on the basis of the known protein structures. Another example is the determination of three-dimensional structures of small proteins and nucleic acids from nuclear magnetic resonance (NMR) NOESY experiments (Wüthrich, 1986). The

0108-7673/90/010046-12$03.00