Acta Cryst. (1996). D52, 1114–1118

Direct-Space Methods in Phase Extension and Phase Refinement. V. The Histogram Moments Method

YUAN-XIN GU, M. M. WOOLFSON AND JIA-XING YAO

Physics Department, University of York, Heslington, York YO1 5DD, England. E-mail: mmw1@york.ac.uk

(Received 29 March 1996; accepted 25 June 1996)

Abstract

Any distribution is completely defined by its moments. It is shown that a process of phase refinement can be carried out, based on Fourier transforms, which modifies the moments of electron density, separately in the protein and solvent regions, towards target values. Tests have been carried out on two moderate-sized proteins with 800–900 atoms in the asymmetric unit, one containing heavy atoms and the other not. It has been found that refinement using the third moment about zero in the protein region is most effective and that refinement with higher moments, or in the solvent region, adds nothing useful. Two kinds of weights are necessary in the method. One is for giving a weighted mixture of new phase indications with original phase estimates from, say, multiple isomorphous replacement. The other weights are applied to the Fourier coefficients of density maps to give the best possible signal:noise ratio. These weights have been explored empirically and the best ones found are described. It is concluded that since the moments method, which changes phases in reciprocal space, is independent of other histogram-matching procedures, which change density in real space, it has something to offer in a refinement package containing several procedures.

1. Introduction

Histogram matching has proved to be a powerful process in phase refinement and extension for proteins. Proteins containing a particular proportion of solvent are quite well characterized by the associated distribution of electron density and the simple process of systematically modifying density, ρ, found with current phase estimates, to density, ρ', which gives a target distribution is found to be quite effective, especially when combined with other refinement criteria as in the *SQUASH* procedure (Zhang & Main, 1990*a*,*b*; Cowtan & Main, 1993). Even more effective are the double-histogram matching techniques described by Refaat, Tate & Woolfson (1996*a*) where the transformed density, ρ', depends not only on ρ, the current density at the grid point in question, but also on some characteristic of the density in a spherical region surrounding the grid point.

All histogram-matching techniques described thus far have depended on directly changing the density in a calculated map. Here, we shall be describing a different approach in which the change of density produces a better distribution for the whole map taken together without considering it in a point-by-point way.

2. Moments and phases

It is known that any distribution is completely defined if all its moments are known and these moments can be about any value, for example, the mean of zero. In general the number of moments will be infinite but if a finite number of moments are known, say t, then it is possible to define a t-block histogram which approximately describes the distribution.

We consider a situation where a trial set of phases is available and from the resultant map it has been possible to find the protein envelope and so to divide the unit cell into protein and solvent regions. The solvent region is expected to be fairly flat and the protein region to have a density histogram which can be derived from a model structure, which could be a known real structure with similar characteristics to the one under investigation.

The electron density is given by,

$$\rho(\mathbf{r}) = \frac{2}{V}\sum_{\mathbf{l}} |F(\mathbf{l})| \cos[2\pi\mathbf{l}\cdot\mathbf{r} - \varphi(\mathbf{l})], \qquad (1)$$

where it is assumed that the space group is $P1$ and that terms of index $\mathbf{l}$ and $\bar{\mathbf{l}}$ have been combined. For brevity writing $\rho = \rho(\mathbf{r})$, we now define the nth moment about zero of the density in the protein region as,

$$A_p^n = \overline{\rho_p^n} = \frac{1}{V_p}\int_p \rho^n \, \mathrm{d}V, \qquad (2)$$

where the integral is taken only over the protein region. Partially differentiating (2) with respect to $\varphi(\mathbf{h})$ gives,

$$\frac{\partial A_p^n}{\partial\varphi(\mathbf{h})} = \frac{1}{V_p}\int_p n\rho^{n-1}\frac{\partial\rho}{\partial\varphi(\mathbf{h})}\,\mathrm{d}V$$

$$= \frac{2n|F(\mathbf{h})|}{VV_p}\int_p \rho^{n-1}\sin[2\pi\mathbf{h}\cdot\mathbf{r} - \varphi(\mathbf{h})]\,\mathrm{d}V. \qquad (3)$$

Writing,

$$X_m(\mathbf{h}) = \frac{1}{V_p}\int_p \rho^m \exp(2\pi i\mathbf{h}\cdot\mathbf{r})\,dV = |X_m(\mathbf{h})|\exp\{i\Psi_p^m(\mathbf{h})\}, \quad (4)$$

we find that,

$$\frac{\partial A_p^n}{\partial\varphi(\mathbf{h})} = \frac{2n|F(\mathbf{h})X_{n-1}(\mathbf{h})|}{V}\sin[\Psi_p^{n-1}(\mathbf{h}) - \varphi(\mathbf{h})]. \quad (5)$$

Similarly, for the solvent region,

$$\frac{\partial A_s^n}{\partial\varphi(\mathbf{h})} = \frac{2n|F(\mathbf{h})Y_{n-1}(\mathbf{h})|}{V}\sin[\Psi_s^{n-1}(\mathbf{h}) - \varphi(\mathbf{h})]. \quad (6)$$

where the subscript s refers to the solvent region and,

$$Y_m(\mathbf{h}) = \frac{1}{V_s}\int_s \rho^m \exp(2\pi i\mathbf{h}\cdot\mathbf{r})\,dV = |Y_m(\mathbf{h})|\exp[i\Psi_s^m(\mathbf{h})]. \quad (7)$$

We now define two residuals,

$$R_p^n = A_{p,0}^n - A_p^n, \quad (8a)$$

and,

$$R_s^n = A_{s,0}^n - A_s^n, \quad (8b)$$

where $A_{p,0}^n$ and $A_{s,0}^n$ are the target values for A_p^n and A_s^n, respectively. A steepest descents method is now used to find the shifts in phases which will eliminate, or reduce, the residuals given by (8*a*) and (8*b*). For a particular n and $\mathbf{h}$ the usual gradient method gives,

$$\Delta\varphi_p^n\,\varphi(\mathbf{h}) = R_p^n \frac{\partial A_p^n}{\partial\varphi(\mathbf{h})} \Big/ \sum_{\mathbf{h}}\left[\frac{\partial A_p^n}{\partial\varphi(\mathbf{h})}\right]^2, \quad (9)$$

for the protein and a similar expression, with subscript s instead of p for the solvent. In test applications we have made so far $n = 2$ to 6 has been used for the protein region but for the solvent region, which is expected to be flat, we have only tried $n = 2$ and 3. The value $n = 1$ is not very informative since $X_0(\mathbf{h})$ and $Y_0(\mathbf{h})$ then depend only on the regions of the cell occupied by protein and solvent and not on the variation of density within those regions. It should also be noted that in an ideal situation where the solvent is completely uniform a value of $\overline{\rho^2}^s = (\bar{\rho}^s)^2$ would give zero variance. There is no real need to consider values of n higher than 2 to describe ideal conditions in the solvent region but, nevertheless, we did consider $n = 3$ just in case we found a useful result.

The refinement strategy given by (9) is satisfactory for general reflections but for special reflections another approach has been used. Here the phase is kept fixed but the magnitude of the structure factor is refined. Now we write,

$$\frac{\partial A_p^n}{\partial F(\mathbf{h})} = \frac{1}{V_p}\int_p n\rho^{n-1}\frac{\partial\rho}{\partial F(\mathbf{h})}\,dV = \frac{2n}{VV_p}\int_p \rho^{n-1}\cos[2\pi\mathbf{h}\cdot\mathbf{r} - \varphi(\mathbf{h})]\,dV = \frac{2}{V}|X_{n-1}(\mathbf{h})|\cos[\varphi(\mathbf{h}) - \Psi_p^{n-1}(\mathbf{h})]. \quad (10)$$

Equation (9) would be applicable where there are only general reflections. Where both general and special reflections occur then the shift in phase or magnitude should be given by,

$$\Delta_p^n q(\mathbf{h}) = R_p^n \frac{\partial A_p^n}{\partial q(\mathbf{h})} \Big/ \sum_{\mathbf{h}}\left[\frac{\partial A_p^n}{\partial q(\mathbf{h})}\right]^2, \quad (11)$$

where $q(\mathbf{h}) = \varphi(\mathbf{h})$ for a general reflection and $F(\mathbf{h})$ for a special reflection.

The phase-changing strategy for special reflections is that at any stage we have available a current weight $w(\mathbf{h})$ which starts with the value unity. In each refinement cycle if

$$w(\mathbf{h})|F(\mathbf{h})| + \Delta F(\mathbf{h}) > 0 \text{ the phase is unchanged}$$

but if $w(\mathbf{h})|F(\mathbf{h})| + \Delta F(\mathbf{h}) < 0$ the phase is changed by π.

The new weight is then given by,

$$w(\mathbf{h})_{\text{new}} = \min\left\{\frac{|w(\mathbf{h})|F(\mathbf{h})| + \Delta F(\mathbf{h})|}{|F(\mathbf{h})|}, 1.0\right\}. \quad (12)$$

This weighting system allows a gradualistic approach towards switching from one special phase to the alternative value.

3. Weighting schemes

Experience with different methods of phase refinement for proteins indicates that good weighting procedures at all stages are very important in getting the best results. Where multiple-isomorphous-replacement (MIR) phase estimates are available then it is better to proceed with a weighted mixture of the phase given by the refinement and the MIR phases. It is also desirable to have other weights indicating the overall reliability of the individual phase estimates which can be applied to the Fourier coefficients of the calculated density maps to increase the map quality in terms of signal:noise ratio and to give for the final map a higher value of the conventional map-correlation coefficient (MCC). The moments technique we report here does not suggest an obvious weighting scheme, as do some methods, so we have tried various schemes based on intuitive ideas. The weighting schemes *B* and *C* described below are the two which have been most effective of the eight different weighting schemes that were investigated. We also give scheme *A* which is reasonably effective without any weights being applied.

3.1. *Scheme A*

This used no weights either for retaining the influence of the MIR phases or to modify the Fourier coefficients when maps were calculated.

3.2. *Scheme B*

The mixing of the phase given by the refinement and MIR phases was effected by,

$$\tan(\varphi_{\text{new}}) = \frac{\eta w_{\text{MIR}} \sin(\varphi_{\text{MIR}}) + E(0)X^{n-1}\sin(\varphi_{\text{old}} + \Delta\varphi)}{\eta w_{\text{MIR}} \cos(\varphi_{\text{MIR}}) + E(0)X^{n-1}\cos(\varphi_{\text{old}} + \Delta\varphi)} = \frac{T}{B}, \tag{12}$$

where η is an adjustable parameter and $\Delta\varphi$ is the shift indicated by the gradient method. The corresponding weight used for calculating the maps at each stage and for the final density map is,

$$w_{\text{map}} = \frac{(T^2 + B^2)^{1/2}}{1 + \eta}, \tag{13}$$

with cut offs at a lower limit of 0.1 and an upper limit of 1.

3.3. *Scheme C*

The mixing of the phase given by the refinement and MIR phases was effected by,

$$\tan(\varphi_{\text{new}}) = \frac{\eta w_{\text{MIR}} \sin(\varphi_{\text{MIR}}) + \frac{1}{2}(1-\eta)E(0)X^{n-1}[1+\cos(\delta\varphi)]\sin(\varphi_{\text{old}} + \Delta\varphi)}{\eta w_{\text{MIR}} \cos(\varphi_{\text{MIR}}) + \frac{1}{2}(1-\eta)E(0)X^{n-1}[1+\cos(\delta\varphi)]\cos(\varphi_{\text{old}} + \Delta\varphi)} = \frac{T}{B}. \tag{14}$$

Again η is an adjustable parameter and $\delta\varphi = \varphi_{\text{old}} + \Delta\varphi - \varphi_{\text{MIR}}$. The weight for calculating maps is $w_{\text{map}} = (T^2 + B^2)^{1/2}$ with lower and upper cut offs of 0.1 and 1.

Although the above analysis has been given in terms of structure amplitudes $|F|$ we have found empirically that the best results are obtained by the use of normalized structure amplitudes $|E|$. This is consistent with our general experience with many processes of phase extension and refinement that have been investigated in this laboratory. Tests with observed $|F|$'s and $|F|$'s with all degrees of sharpening indicate that the use of $|E|$ usually gives the best results.

4. Tests of the refining algorithm

It must be said at this stage that although the theory we have given, which was the basis of our tests, is perfectly valid the actual procedures which turned out to be the most effective did not use all the fine detail of the theory. Thus, when (11) was applied to give shifts some of the changes were unrealistically large - bigger than 360 , for example. The reason for this was that the linear theory was being taken well beyond its range of applicability. To deal with this problem phase shifts were chosen by a much simpler formula,

$$\Delta\varphi(\mathbf{h}) = k\frac{\partial A_p^n}{\partial\varphi(\mathbf{h})}, \tag{15}$$

where k was taken to make the mean phase shift equal to some value chosen by experience. For special reflections we found that (11) could still be used.

Most of the tests which have been made are for $n = 3$. Tests for higher values of n gave results which were highly correlated with the $n = 3$ results, slightly worse for $n = 4$ and much worse for $n = 5$ and $n = 6$. The correlation between the results for different n can be understood intuitively. Any change of phase which alters the distribution of the values of ρ in such a way that there is a bias towards larger values will increase all the moments and conversely if the bias is in the other direction the moments will all reduce in value. No good way of combining the results for different values of n has been found since all combinations tried gave results worse than for $n = 3$ alone.

Values of n equal to 1 or 2 are also not very useful. The values of $\bar{\rho}$ and $\overline{\rho^2}$, where the averages are taken over the whole cell, are structure-invariant quantities and, since most of the density resides in the protein region, the averages over the protein region are not very sensitive to phase variations. The same situation exists with shifts given by the moments in the solvent region. The results for the solvent region alone were found to be much worse than for $n = 3$ in the protein region and no fruitful way of combining the results could be found.

Table 1. *Refinement of 6450 1.9 Å MIR phases of 2-Zn insulin*

Initial mean phase error (MPE) was 62.1° with MCC = 0.372. NC = number of refinement cycles; MPE = unweighted mean phase error; MCC = MCC for map with unweighted Fourier coefficients; WMPE = weighted MPE with weights w_{map}; WMCC = MCC for map with Fourier coefficients $w_{\text{map}}E$; EMPE = E-weighted MPE; WEMPE = w_{map} E-weighted MPE.

	Weighting scheme		
	A	*B* ($\eta = 0.1$)	*C* ($\eta = 0.1$)
NC	3	3	3
MPE (°)	53.4	54.5	54.3
MCC	0.504	0.512	0.515
WMPE (°)		46.4	46.5
WMCC		0.572	0.570
EMPE (°)	49.9	50.6	50.4
WEMPE (°)		43.5	43.7

For all the reasons given the results reported here are only for $n = 3$ in the protein region.

The first test we report is for MIR phases at 1.9 Å resolution for 2-Zn insulin (Baker, Blundell, Cutfield, Cutfield, Dodson, Dodson, Hodgkin, Hubbard, Isaacs, Reynolds, Sakabe, Sakabe & Vijayan, 1988). The space group is $R3$ with $a = 82.5$, $c = 34.0$ Å, $Z = 9$. The asymmetric unit contains 831 non-H atoms including two Zn atoms but excluding solvent. The results of refinement are shown in Table 1.

In these trials the mean phase shift per cycle, as controlled by the parameter k in (15), were 20, 20 and 10 in the three cycles used as this pattern has been found to be satisfactory.

A trial was also made with a structure of similar size to 2-Zn insulin RNAp1 (Bezborodova, Ermekbaeva, Shlyapnikov, Polyakov & Bezborodov, 1988) but containing no atoms heavier than S atoms. Space group $P2_1$ with $a = 32.01$, $b = 43.76$, $c = 30.67$ Å, $\beta = 115.83$, $Z = 2$. The asymmetric unit contains 808 non-H atoms in the asymmetric unit, including five S atoms, plus 83 water molecules. There are 23 853 independent reflections to 1.17 Å resolution. In this case errors were artificially imposed on the calculated phases to give an initial mean phase error of 65.4 but the observed data were used in the refinement process. After two cycles using scheme *A* (no weights) the MPE had dropped to 55.9 .

5. Comments and conclusions

The tests reported here are a small sample of many tests which have been carried out with various schemes for weighting the Fourier maps which are required and for combining new phases estimated with the MIR phases. In general the efficiency of the moments method for phase refinement is similar to, but slightly worse than, that of the normal histogram-matching method but considerably less than that of either of the double-histogram matching methods (Refaat, Tate & Woolfson, 1996*a*).

It is possible that an even more exhaustive investigation of the moments method would lead to some improvements but we cannot see that it would ever approach the effectiveness of, say, the double-histogram method or the LDE (low-density elimination) method (Shiono & Woolfson, 1991; Refaat & Woolfson, 1993). However, its main strength lies in its independent approach, modifying phases in reciprocal space rather than density in real space, so that used in conjunction with other histogram-matching methods it may have something extra to offer. For example, a combination involving nine cycles of the moments method interspersed with three cycles of the double-histogram-with-local-variance method gave a map with WMCC = 0.610 for the 2-Zn insulin 1.9 Å data.

It is worth pointing out that a new procedure, even if less effective than some previous procedures, may still have something to offer as long as it is independent of those previous procedures. A single step of the moments method gives a quite different pattern of change of density from that given by a single step of normal histogram matching, for example, since there is no strong correlation between the change of density during the step and the original density. To illustrate this principle from the field of statistics, if several independent measurements of the same quantity, X, are made with different variances then an inverse-variance-weighted average of them all gives a combined estimate with a lower variance (uncertainty) than any of the individual measurements, including the best one. This general principle is worth bearing in mind. The phase extension and refinement system *SQUASH* contains components of very different refinement capability but the combination is more powerful than any of them used alone. The moments method with $n = 3$ is to be incorporated in the phase extension and refinement program, *PERP* (Refaat, Tate & Woolfson, 1996*b*) which contains seven other phase refinement procedures, all of different capabilities but together more effective than any of them.

It will be clear that, despite the introduction of this method as one involving moments, when it comes to application it is only the third moment which has been exploited. It seems plausible that the normal histogram-matching method, which is matching all moments simultaneously, is bringing more information to bear and therefore should be the more powerful method. On the other hand our results indicate that the information from higher moments is heavily correlated with the third-moment information so that, perhaps, not much more is being obtained from the full histogram than is being obtained from the third moment alone.

We are grateful to the Academia Sinica, the Royal Society and the Leverhulme Trust for support which has made possible this research and other collaboration between the Institute of Physics, Beijing, and the Physics Department of the University of York.

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.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148–157.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1996*a*). *Acta Cryst.* D**52**, 252–256.
Refaat, L. S., Tate, C. & Woolfson, M. M. (1996*b*). *Acta Cryst.* D**52**.
Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.
Shiono, M. & Woolfson, M. M. (1991). *Acta Cryst.* A**47**, 526–533.
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.