Acta Cryst. (1993). D**49**, 530–533

Mean Phase Error and the Map Correlation Coefficient

BY V. Y. LUNIN

Institute of Mathematical Problems of Biology, Moscow Region, Pushchino 142292, *Russia*

AND M. M. WOOLFSON

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

(*Received* 26 *April* 1993; *accepted* 9 *June* 1993)

Abstract

In judging the effectiveness of methods of solving crystal structures, or in phase refinement and development, two criteria are commonly used. The first is the mean phase error, which may be weighted in some way, and the second is the map correlation coefficient which describes the similarity of a map with estimated phases to that with true phases. It is shown that these two measures are directly related and that given the individual phase errors the map correlation coefficient may be found without the need to calculate a map. Various aspects of this connection are examined, including the map correlation coefficient when weights are used for calculating maps and the conditions under which phase extension leads to maps with a higher map correlation coefficient – which involves a balance between the advantage of employing more data and the disadvantage that the extra data may have a higher average phase error.

Introduction

In the development of new methods of solving crystal structures or of phase extension and refinement it is customary to use known structures for testing purposes. By this means the actual effectiveness of procedures can be assessed by comparison of the results obtained with some target – for example, the phases calculated from the finally refined model, which can be thought of as *true* phases.

In general those working on the development of direct methods have tended to express their results in terms of phase errors

$$\Delta\varphi(\mathbf{h}) = \varphi_t(\mathbf{h}) - \varphi_e(\mathbf{h}) \qquad (1)$$

where $\varphi_t(\mathbf{h})$ and $\varphi_e(\mathbf{h})$ are the true and estimated phases, respectively. The most commonly quoted overall measures of effectiveness are either the mean phase error $\overline{|\Delta\varphi|}$ or the root-mean-square phase error $\overline{(\Delta\varphi)^2}^{1/2}$. Very often the process of phase determination or phase refinement gives phase errors correlated with the structure amplitude and in this case weighted phase errors are informative with weights $|E|$ commonly used, where the E's are the normalized structure factors.

In the field of protein crystallography where, in general, the data do not extend to atomic resolution ($\sim$1.5 Å), the critical stage in structure determination is the production of a map which can be interpreted in terms of some molecular model. Silva & Viterbo (1980) considered the relationship between the root-mean-square phase error and the quality of the map calculated with estimated phases. They restricted their analysis to small structures with data at atomic resolution and the map quality was judged by the number of map peaks within some reasonably small distance of the atomic positions. Such an approach is clearly inappropriate to low-resolution data since atomic peaks would not occur for such comparisons of distance to be made. A figure of merit based on the probable interpretability of a map is regarded as the most relevant in developing methods to be applied to proteins and this is often taken as the map correlation coefficient, defined as

$$R = \frac{\overline{\rho_t(\mathbf{r})\rho_e(\mathbf{r})} - \overline{\rho_t(\mathbf{r})}\ \overline{\rho_e(\mathbf{r})}}{[\overline{\rho_t^2(\mathbf{r})} - \overline{\rho_t(\mathbf{r})}^2]^{1/2}[\overline{\rho_e^2(\mathbf{r})} - \overline{\rho_e(\mathbf{r})}^2]^{1/2}}, \qquad (2)$$

where $\rho_t(\mathbf{r})$ and $\rho_e(\mathbf{r})$ are the map densities found with the true and estimated phases and the averages are over the whole unit cell. The values of density used for the averages are conventionally taken at grid points in a cell and the value of R is then the normal linear correlation coefficient of the two sets of numbers so obtained. A value of R greater than 0.5 usually indicates a promising starting point for map interpretation but sometimes successful structure determinations can be started with substantially lower values. Refaat & Woolfson (1993) showed this by the application of the low-density elimination (LDE) method to the known structure of ribonuclease RNAp1 (Bezborodova, Ermekbaeva, Shlyapnikov, Polyakov & Bezborodov, 1988). This structure has space group $P2_1$ with $a = 32.01$, $b =$

ISSN 0907-4449 ©1993

49.76, $c = 30.67$ Å and $\beta = 115.83°$ and contains 96 amino-acid residues plus 83 water molecules in the asymmetric unit. There are 28853 independent observed reflections out to 1.17 Å resolution. By means of the LDE method, which involves no intermediate interpretation of the map, the map correlation coefficient was increased from a starting value of 0.222 to a final value of 0.697. Since ribonuclease RNAp1 is a small protein with high-resolution data it may be unwise to generalize from it, so we suggest that 0.4 may be a more realistic estimate of a potentially useful starting point.

It is clear that there is some relationship between mean phase errors and map correlation coefficients and our purpose here is to show the form of that relationship.

Mathematical analysis

The relationship between true density and structure factors is given by

$$\rho_t(\mathbf{r}) = (1/V)\sum_{\mathbf{h}} F_t(\mathbf{h})\exp(-2\pi i\mathbf{h}.\mathbf{r}), \qquad (3)$$

where $F_t(\mathbf{h})$ has the observed magnitude, $|F_o(\mathbf{h})|$, and the true phase, $\varphi_t(\mathbf{h})$. Since $\exp(-2\pi i\mathbf{h}.\mathbf{r})$ averaged over the unit cell is zero, unless $\mathbf{h} = 0$ when it equals unity, then it is evident that

$$\overline{\rho_t(\mathbf{r})} = (1/V)F(0) \qquad (4)$$
$$= \overline{\rho_e(\mathbf{r})},$$

since it is phase independent. From (3) we also find

$$\rho_t^2(\mathbf{r}) = (1/V^2)\sum_{\mathbf{h}}\sum_{\mathbf{k}} F_t(\mathbf{h})F_t(\mathbf{k})\exp[-2\pi i(\mathbf{h}+\mathbf{k}).\mathbf{r}]. \qquad (5)$$

When averages are taken on the two sides the only finite contributions on the right-hand side are when $\mathbf{h} + \mathbf{k} = 0$ or $\mathbf{k} = -\mathbf{h}$ so that

$$\overline{\rho_t^2(\mathbf{r})} = (1/V^2)\sum_{\mathbf{h}} |F_o(\mathbf{h})|^2 \qquad (6)$$
$$= \overline{\rho_e^2(\mathbf{r})},$$

since it is phase independent. In a similar way to (5) we find

$$\rho_t(\mathbf{r})\rho_e(\mathbf{r}) = (1/V^2)\sum_{\mathbf{h}}\sum_{\mathbf{k}} F_t(\mathbf{h})F_e(\mathbf{k})\exp[-2\pi i(\mathbf{h}+\mathbf{k}).\mathbf{r}], \qquad (7)$$

and after averaging the two sides

$$\overline{\rho_t(\mathbf{r})\rho_e(\mathbf{r})} = (1/V^2)\sum_{\mathbf{h}} F_t(\mathbf{h})F_e(\bar{\mathbf{h}}). \qquad (8)$$

After combining terms for which the pairs of indices are $\mathbf{h}$, $\bar{\mathbf{h}}$ and $\bar{\mathbf{h}}$, $\mathbf{h}$ we find

$$\overline{\rho_t(\mathbf{r})\rho_e(\mathbf{r})} = (1/V^2)\sum_{\mathbf{h}} |F_o(\mathbf{h})|^2\cos[\Delta\varphi(\mathbf{h})]. \qquad (9)$$

Inserting results (2), (6) and (9) into (2) we find

$$R = \sum_{\mathbf{h}(\mathbf{h}\neq 0)} |F_o(\mathbf{h})|^2\cos[\Delta\varphi(\mathbf{h})] \Big/ \sum_{\mathbf{h}(\mathbf{h}\neq 0)} |F_o(\mathbf{h})|^2. \qquad (10)$$

This shows that R is an $|F|^2$-weighted average of the values of $\cos(\Delta\varphi)$, with the origin term, which has zero mean phase error, excluded. To allow for symmetry, the summations in (10) should be made over the whole of reciprocal space; if they are only made over one asymmetric unit of reciprocal space then appropriate multiplicity factors will need to be associated with the structure factors, according to their type. It should also be noted that (10) is an exact result; if the right-hand side is calculated using the individual phase errors then the map correlation coefficient is available, a measure of the quality of the resultant density map. For estimated phases equal to the true phases the value of R is 1.0 while for random phases it will be zero.

Use of modified maps

In direct-methods work it is usual to calculate E maps rather than F maps because, with data at atomic resolution they are better at defining the position of an atomic peak, albeit that they introduce noise in the form of diffraction ripples in other parts of the map. Equation (10) requires no modification in this case; E's are used in place of F's and the correlation is found with respect to an E map with perfect phases.

What is not so straightforward is when a weighting scheme is used which modifies the magnitude of the Fourier coefficient according to an assessment of the likely error of the phase estimate. For example in the low-density elimination scheme described by Shiono & Woolfson (1992), which involved the iterative modification of density maps, a weighted E map was used where the weight was

$$w(\mathbf{h}) = \tanh[\mathcal{F}(\mathbf{h})E(\mathbf{h})/2\overline{\mathcal{F}^2}^{1/2}] \qquad (11)$$

where $\mathcal{F}(\mathbf{h})$ is the magnitude of the $\mathbf{h}$th Fourier coefficient of the previous map and $\overline{\mathcal{F}^2}^{1/2}$ is the root-mean-square value of the $\mathcal{F}$'s. Given that the normal E map, or F map, with true phases is the ideal target of the phasing procedure then the value of R should be calculated with respect to this. By similar reasoning which gave rise to (10) it can be shown that where a weight $w(\mathbf{h})$ is associated with $F(\mathbf{h})$ then the corresponding map correlation coefficient is given by

$$R_w = \frac{\sum_{\mathbf{h}(\mathbf{h}\neq 0)} w(\mathbf{h})|F_o(\mathbf{h})|^2\cos[\Delta\varphi(\mathbf{h})]}{\left[\sum_{\mathbf{h}(\mathbf{h}\neq 0)} |F_o(\mathbf{h})|^2\right]^{1/2}\left[\sum_{\mathbf{h}(\mathbf{h}\neq 0)} w(\mathbf{h})^2|F_o(\mathbf{h})|^2\right]^{1/2}}. \qquad (12)$$

Given that the values of $\Delta\varphi(\mathbf{h})$ are known this is easily calculated.

Use of partial phase information

Let $\mathscr{S}$ be the whole set of reflections for which the structure amplitudes $|F(\mathbf{h})|$ or $|E(\mathbf{h})|$ are known. Frequently phases are determined for only a subset, $\mathscr{S}_1$, of the reflections, for example reflections at low resolution or a subset or large normalized structure factors. For such a situation, by reasoning similar to that which gave (10) and (12), we find the initial value of R

$$R_{\text{init}} = \frac{\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})]}{\left[\sum_{\mathbf{h}\in\mathscr{S}} |F_o(\mathbf{h})|^2\right]^{1/2}\left[\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2\right]^{1/2}}. \quad (13)$$

The next stage is to find some additional phases for some set of reflections, $\mathscr{S}_2$ and the map correlation coefficient for the extended phases becomes

$$R_{\text{ext}} = \frac{\sum_{\mathbf{h}\in\mathscr{S}_1\cup\mathscr{S}_2} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})]}{\left[\sum_{\mathbf{h}\in\mathscr{S}} |F_o(\mathbf{h})|^2\right]^{1/2}\left[\sum_{\mathbf{h}\in\mathscr{S}_1\cup\mathscr{S}_2} |F_o(\mathbf{h})|^2\right]^{1/2}}. \quad (14)$$

The question which now arises is whether the extended phases give a better, *i.e.* higher, value of R than the initial ones; clearly if the extended phases were all absolutely correct the map would be better but if they were random then the map would be worse. The condition for a map with a larger map correlation coefficient is

$$R_{\text{ext}} > R_{\text{init}} \quad (15)$$

or, in the full form

$$\frac{\sum_{\mathbf{h}\in\mathscr{S}_1\cup\mathscr{S}_2} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})]}{\left[\sum_{\mathbf{h}\in\mathscr{S}} |F_o(\mathbf{h})|^2\right]^{1/2}\left[\sum_{\mathbf{h}\in\mathscr{S}_1\cup\mathscr{S}_2} |F_o(\mathbf{h})|^2\right]^{1/2}} > \frac{\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})]}{\left[\sum_{\mathbf{h}\in\mathscr{S}} |F_o(\mathbf{h})|^2\right]^{1/2}\left[\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2\right]^{1/2}}. \quad (16)$$

This inequality may be transformed to

$$\frac{\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})] + \sum_{\mathbf{h}\in\mathscr{S}_2} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})]}{\left[\sum_{\mathbf{h}\in\mathscr{S}_1\cup\mathscr{S}_2} |F_o(\mathbf{h})|^2\right]^{1/2}} > \frac{\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})]}{\left[\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2\right]^{1/2}}, \quad (17)$$

or

$$\sum_{\mathbf{h}\in\mathscr{S}_2} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})] > \sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})] \left\{ \frac{\left[\sum_{\mathbf{h}\in\mathscr{S}_1\cup\mathscr{S}_2} |F_o(\mathbf{h})|^2\right]^{1/2}}{\left[\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2\right]^{1/2}} - 1 \right\} \quad (18)$$

If we now write as the *inner* correlation coefficient of a set $\mathscr{S}_k$

$${}_kR_{in} = \left\{\sum_{\mathbf{h}\in\mathscr{S}_k} |F_o(\mathbf{h})|^2 \cos[\Delta\varphi(\mathbf{h})]\right\} \Big/ \left[\sum_{\mathbf{h}\in\mathscr{S}_k} |F_o(\mathbf{h})|^2\right] \quad (19)$$

and write

$$q = 2\frac{\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2}{\sum_{\mathbf{h}\in\mathscr{S}_2} |F_o(\mathbf{h})|^2} \left\{ \left[1 + \frac{\sum_{\mathbf{h}\in\mathscr{S}_2} |F_o(\mathbf{h})|^2}{\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2} \right]^{1/2} - 1 \right\}, \quad (20)$$

then condition (15) that the map correlation coefficient should increase becomes

$${}_2R_{\text{in}} > \tfrac{1}{2}q_1R_{\text{in}}. \quad (21)$$

The factor q may be expressed as a function of the ratio

$$t = \left[\sum_{\mathbf{h}\in\mathscr{S}_2} |F_o(\mathbf{h})|^2\right] \Big/ \left[\sum_{\mathbf{h}\in\mathscr{S}_1} |F_o(\mathbf{h})|^2\right], \quad (22)$$

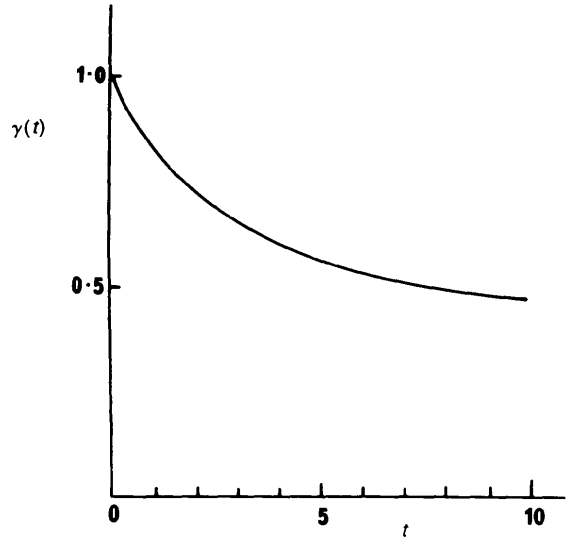

Fig. 1. The function $\gamma(t)$. The quantity t is the ratio of the sum of the squares of the structure amplitudes of the extension set to that of the original set of reflections. $\gamma(t)$ defines the condition under which the extension set will increase the map correlation coefficient.

and is equal to

$$q = 2[(1+t)^{1/2} - 1]/t = \gamma(t). \qquad (23)$$

The function $\gamma(t)$ is shown in Fig. 1; it equals 1 for $t = 0$ and decreases monotonically with increasing t. Since $\gamma(t) \leq 1$ for all $t \geq 0$ then it follows that ${}_2R_{\rm in} > \frac{1}{2}{}_1R_{\rm in}$ is a sufficient condition for $R_{\rm ext} > R_{\rm init}$.

If the value of t, as defined by (22) is small then q is very close to 1 so that the condition for improving the map correlation coefficient by estimating more phases is that the inner correlation coefficient for the extension phases must be at least half of that for the starting phase set. In a particular case if the extension set $\mathcal{E}_2$ contains only one phase and all the phases of the starting set $\mathcal{E}_1$ are correctly determined then condition (21), which applies to the single reflection in $\mathcal{E}_2$, becomes

$$\cos(\Delta\varphi) > \tfrac{1}{2}$$

or

$$|\Delta\varphi| < 60°. \qquad (24)$$

It should also be noted that since ${}_1R_{\rm in} \leq 1$ and $q \leq 1$ the condition

$${}_2R_{\rm in} > \tfrac{1}{2}$$

is sufficient to have improved correlation. This implies that if the $|F|^2$-weighted mean of $\cos\Delta\varphi$ of the extension set is greater than 0.5 (corresponding to $|\Delta\varphi| < 60°$ if all the phase errors had the same magnitude) then the extended map would have a higher correlation coefficient whatever the phase errors of the original phase set. As a more general example we consider ${}_1R_{\rm in}$ equal to 0.7, corresponding to a mean phase error of order 45°; if $q = 0.8$, implying $t \simeq 1$, so that the sum of intensities of the extension set equals that of the original set, then (21) implies that ${}_2R_{\rm in} > 0.28$. In such a case mean phase errors of the extension set, even as high as 73°, will increase the map correlation coefficient.

We are grateful to the Royal Society for the award of a Kapitza Fellowship to one of us (VYL) during the tenure of which this work was carried out. We are also grateful to useful comments by referees which led to improvements of presentation.

References

Bezborodova, S. I., Ermekbaeva, L. A., Shlyapnikov, S. V., Polyakov, K. M. & Bezborodov, A. M. (1988). *Biokhimiya*, **53**, 965–968.

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

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

Silva, A. M. & Viterbo, D. (1980). *Acta Cryst.* A**36**, 1065–1070.