Acta Cryst. (1994). **D50**, 842–846

On the Application of Phase Relationships to Complex Structures. XXXIV. VFOM – a New Figure of Merit for Protein Phase Sets at Moderate Resolution

By A. F. Mishnev* and M. M. Woolfson

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

(*Received* 3 *February* 1994; *accepted* 2 *June* 1994)

Abstract

In recent years it has been shown that direct methods are capable of solving the structures of small proteins. Mukherjee & Woolfson [*Acta Cryst.* (1993), **D49**, 9–12] have shown that useful phase sets can be produced even at 3 Å resolution but that the standard figures of merit could not distinguish the better phase sets from others. They found modified forms of the standard figures of merit that could pick out better phase sets for 2 Å resolution or higher. Gilmore, Henderson & Bricogne [*Acta Cryst.* (1991), **A47**, 842–846] have shown that evaluation of the log-likelihood gain, coming from entropy-maximization procedures, is also very successful in picking out good protein phases sets. A new figure of merit is described, based on the expected charactistics of an electron-density map for a protein, and comparisons are made with the other figures of merit mentioned above.

Introduction

When a set of protein phases is available from a method based on isomorphous replacement, anomalous scattering or a combination of the two then the crystallographer can try to interpret the resultant map with reasonable confidence that it has some meaningful content. In the event that an interpretation is unsuccessful then a phase extension and refinement process, such as *SQUASH* (Zhang & Main, 1990*a*,*b*; Cowtan & Main, 1993), will often so improve the map that interpretation becomes possible.

The situation with direct-methods-generated phases is very different. Typically 1000 sets of phases will be generated and if no figures of merit (FOM's) are available to distinguish the better sets from the poorer ones then nothing can be done with them. It is impracticable to examine each of the maps with a view to interpretation or phase refinement and extension. Mukherjee & Woolfson (1993) using the direct-methods procedure *SAYTAN* (Debaerdemaeker, Tate & Woolfson, 1988) found it possible to generate useful sets of phases for the small protein aPP (Glover *et al.*, 1983) even with 3 Å data but the conventional *SAYTAN* FOM's were unable to pick them out. Mukherjee & Woolfson managed to modify the conventional FOM's into a form which did select the better phase sets, but only for resolutions higher than 2 Å. These modified FOM's were still heavily based on statistical principles and the property of a map for a small structure, that $\int \rho^3 dV$ should be maximum, which is not true for larger structures. Since aPP has only 36 amino acids, one Zn atom and 80 water molecules in the asymmetric unit, the conditions applying to small structures are just managing to give some discriminating information but it must be expected that any FOM based on small-structure properties will fail for any structure much larger than this one.

A successful FOM has been described by Gilmore, Henderson & Bricogne (1991) which is based on the evaluation of log-likelihood gain, which comes from entropy maximization. This was tried on *SAYTAN* sets of phases, with data at 0.98 Å resolution, and shown to be capable of recognizing sets of phases with mean phase error (MPE) less than about 50°. No application to lower resolution phase sets has been reported.

Here we describe our derivation of a new FOM which should apply to larger protein structures and lower resolutions.

Theoretical background

One of the first steps in determining a protein structure when an initial map is available is to distinguish that part of the cell occupied by the protein from that occupied by the solvent. One approach is to use the fact that the mean density in the protein region, occupied by a more-or-less rigid structure, is higher than that in solvent region occupied by mobile and less densely packed solvent molecules (Bhat & Blow, 1982; Wang, 1985). The Wang procedure is very widely used in protein crystallography; once the protein envelope has been defined a process of density flattening in the solvent region is found to be effect-

* Permanent address: Latvian Institute of Organic Synthesis, Riga LV 1006, Latvia.

ive for phase extension and refinement. It can also be used to resolve the phase ambiguity when single-isomorphous replacement or one-wavelength anomalous-scattering techniques are used.

Another approach is based on the observation that the density in the protein region is not only higher on average than in the solvent region but that it also has greater variability (Reynolds *et al.*, 1985). Thus, in the protein regions there is found both the highest density, corresponding to atomic peaks, but also the lowest density in regions most remote from atomic centres. It is worth noting that the Wang procedure does contain this condition, albeit in a rather weak form. In the Wang process the local average density is formed but, before that, all negative density in the map is replaced by zero. Since the negative density is going to occur mainly, or perhaps only, in the protein region this has the effect of increasing the contrast between protein and solvent regions when the local averaging is carried out. A technique for finding the protein envelope at 4 Å resolution for the structure of tumour necrosis factor based on the variability of density has been reported by Jones, Walker & Staurt (1991) and it gave a better definition of the protein region than the Wang method did. This structure, with space group $P3_121$, contains six protein units, each with 157 amino acids in the asymmetric unit so it is clear that the characteristics of protein maps being used in this application are valid for structures of this size and with data of this resolution.

This led us to consider a new FOM,

$$\mathrm{VF} = \sum_i \overline{\rho'_i}\,(\overline{\rho'^2_i} - \overline{\rho'_i}^{\,2}), \qquad (1)$$

where ρ' is the map density with negative regions made equal to zero and $\overline{\rho'_i}$ and $\overline{\rho'^2_i}$ are the average values of ρ' and ρ'^2 in a sphere of radius R surrounding the grid point i. The term in parentheses in (1) is the local variance of ρ', V'_i. In calculating VF we used a procedure, also suggested by Leslie (1987), which consists of the following steps.

(i) Calculate an E map with phases from the direct-methods program but with the origin term $E(\mathbf{0})$ removed.

(ii) Replace all negative density by zero to give ρ'. The removal of the origin term $E(\mathbf{0})$ from the map increases the volume of negative density, which will be mostly in the protein region. This increases the contrast in average density between the protein and solvent regions when negative density is replaced by zero.

(iii) By Fourier transformation find $E'(\mathbf{h})$ and $G'(\mathbf{h})$, the Fourier transforms of ρ' and ρ'^2, respectively.

(iv) Multiply $E'(\mathbf{h})$ and $G'(\mathbf{h})$ by the Fourier transform of a sphere of chosen radius R, $Q(\mathbf{h})$.

Table 1. *Values of VF* (1) *for different resolutions of aPP and with different random mean phase errors (MPE's) applied to calculated data*

The upper figures (× 10^2) are obtained from the accurate calculation and the lower (in parentheses) from the approximation using the averages in boxes of size approximately 2 × 2 × 2 Å.

	Resolution (Å)			
MPE (°)	1.0	1.5	2.0	2.5
0	7315	821	270	159
	(3753)	(1501)	(718)	(340)
~ 20	6706	763	251	149
	(3199)	(1303)	(623)	(302)
~ 40	4660	579	197	127
	(1872)	(802)	(406)	(225)
~ 60	3609	487	177	117
	(1326)	(589)	(329)	(192)
~ 80	2972	413	169	110
	(1234)	(552)	(318)	(186)

(v) With $E'(\mathbf{h})$ $Q(\mathbf{h})$ and $G'(\mathbf{h})$ $Q(\mathbf{h})$ as Fourier coefficients calculate $\overline{\rho'}$ and $\overline{\rho'^2}$.

(vi) Using the values of $\overline{\rho'}$ and $\overline{\rho'^2}$ calculated at grid points evaluate VF.

The first test of VF was made with aPP, truncating the data to different resolutions and with random errors with different MPE's imposed on the calculated phases. Trial and error showed that a value of $R = 5$ Å gave reasonable results and this value is used throughout all the tests which we describe. The results are shown in Table 1 and it will be seen that the principle of the VF FOM is sound and that it sharply discriminates in favour of sets of phases with lower MPE's.

The procedure described above requires five fast Fourier transforms (FFT's) but we have also explored a simpler approach in which only one FFT is required. The original density map is divided into parallelepiped-shaped boxes with edges of dimension between 1 and 2 Å and within each box the values of $\overline{\rho'}$ and V' are calculated by direct summation over all the contained grid points. The results of this approximate calculation are also shown in parentheses in Table 1. While the absolute values of VF are smaller, because the summation includes fewer terms, the values still discriminate well in favour of smaller MPE's.

Tests with *SAYTAN* phase sets

In the preliminary tests of VF, described above, we used the complete data set out to the required resolution and applied errors randomly to the phases. The situation with phase sets from a direct-methods approach is quite different; firstly there will only be a subset of large $|E|$'s with estimated phases and secondly the phase errors, $\Delta\varphi(h)$, which come from direct methods tend to be heavily correlated. This comes about because of the well known three-phase relationships,

$$\varphi(\mathbf{h}) - \varphi(\mathbf{k}) - \varphi(\mathbf{h}-\mathbf{k}) \simeq 0. \qquad (2)$$

Phases derived by the use of the tangent formula, whether they are close to being correct or not, tend to obey relationship (2) so that,

$$\{\varphi(\mathbf{h}) + \Delta\varphi(\mathbf{h})\} - \{\varphi(\mathbf{k}) + \Delta\varphi(\mathbf{k})\}$$
$$- \{\varphi(\mathbf{h}-\mathbf{k}) + \Delta\varphi(\mathbf{h}-\mathbf{k})\} \simeq 0. \qquad (3)$$

Subtracting (2) from (3) gives,

$$\Delta\varphi(\mathbf{h}) - \Delta\varphi(\mathbf{k}) - \Delta\varphi(\mathbf{h}-\mathbf{k}) \simeq 0, \qquad (4)$$

which clearly shows the correlation between the phase errors. This correlation will be less strong for proteins where the three-phase relationship does not hold very strongly and *SAYTAN*, which includes components other than the normal tangent formula, also gives a weaker phase-error correlation.

The net effect of the patterns of phases which come from a direct-methods approach is that they all tend to give peaky maps; in a substantially correct map the peaks will show the positions of the atoms but in an incorrect map they can be in other places in the cell. Sometimes they will fall close to symmetry elements and will, therefore, be in regions forbidden by stereochemical considerations – for example, an atom less than 0.7 Å from a twofold axis will be less than 1.4 Å from its symmetry-related partner. To take account of this we decided to strengthen VF by converting it to the form,

$$\mathrm{VF}' = \sum_i \overline{\rho'_i}\, V'_i - \sum_f \overline{\rho'_f}\, V'_f, \qquad (5)$$

where the first sum includes the allowed grid points and the second sum the forbidden ones. This form of FOM was applied to 400 sets of phases derived by *SAYTAN* from 0.98 Å data for aPP. In running *SAYTAN* there were employed 800 large $|E|$'s and 200 small ones. Three of the phase sets had unweighted MPE's in the range 38–40°. Table 2 lists the VF′ values for the three good sets and a selection of others. It can be seen that with high-resolution data and MPE's about 40° good phase sets can readily be identified by VF′ even using the approximate method of density averaging in boxes. The three good sets had the highest values of VF′ and were appreciably better than any others.

Table 2. *Values of VF′* (5) *calculated for good phase sets and a selection of poor phase sets from a SAYTAN run for aPP with data resolution* 0.98 Å

The approximate method of averaging in boxes was used.

Phase set number	Mean phase error (°)	VF′ ($\times 10^{-1}$)
1	84	109
2	84	104
3	81	95
19	39	187
23	82	139
36	84	84
44	80	87
52	78	87
103	85	135
107	84	140
117	83	34
134	83	134
300	40	168
347	38	166

The next step was to repeat the VF′ test on phase sets produced when the aPP data were truncated to 2 Å. In this case *SAYTAN* selected 600 large $|E|$'s and 200 weaker ones and 23 of the 1000 phase sets generated had MPE's in the range 62–71° and were expected to contain useful structural information. We found that calculating VF′ by both the simplified and accurate procedures did not identify the better phase sets. While they did have values of VF′ much higher than the average for the 1000 trials, for many poor phase sets the values were as high or even higher.

It is obvious that the problem of distinguishing phase sets with MPE's in the range 60–70° from those in the range 80–87° is going to be much more difficult than when the good phase sets have an MPE of 40°. To overcome these difficulties we once again considered the properties of maps corresponding to good and poor phase sets. For a perfect set of phases, but for the effect of Fourier series termination, there would be no negative density and we might expect that maps with poor phases have greater negativity than maps with good phases. However, $\bar{\rho}$, the average density over the whole cell, is a structure-invariant quantity and this means that when we replace negative density by zero the resultant average modified density over the whole cell, $\overline{\rho'}$, will tend to be bigger for an incorrect set of phases than for a correct set. This will tend to enhance the value of VF′ for an incorrect phase set. On the other hand the value of $\overline{\rho^2}$ is also a structure invariant quantity and removing negative density in this case makes $\overline{\rho'^2}$ and also V' smaller for a poor set of phases. With these considerations in mind we produced a final normalized form of the FOM,

$$\mathrm{VFOM} = \Big(\sum_i \overline{\rho'_i}\, V'_i - \sum_f \overline{\rho'_f}\, V'_f\Big)\Big/\overline{\rho'}\, V', \qquad (6)$$

where in the divisor the average and the variance are for the whole cell. This expression was used for the 1000 phase sets produced by *SAYTAN* for the 2 Å data of aPP and in Table 3 there is shown the top of the list of values of VFOM in ranking order. It is remarkable that 22 of the 23 best phase sets appear at the top of the list, even those with 71° MPE, although there is poor contrast between values of VFOM for the better phase sets and the rest.

Table 3. *The top of the VFOM* (6) *ranking order for phase sets from SAYTAN for aPP with* 2 Å *data* (*R* = 5 Å)

Rank number	Mean phase error (°)	VFOM ($\times 10^{6}$)
1	62	0.3032
2	63	0.3027
3	64	0.3025
4	63	0.3022
5	63	0.3022
6	63	0.3021
7	63	0.3021
8	66	0.3021
9	67	0.3017
10	66	0.3016
11	66	0.3013
12	66	0.3013
13	66	0.3010
14	71	0.3000
15	63	0.2997
16	66	0.2994
17	63	0.2993
18	64	0.2989
19	63	0.2989
20	71	0.2989
21	82	0.2988
22	66	0.2986
23	64	0.2986
24	85	0.2982
25	74	0.2975
26	82	0.2975
27	84	0.2975
28	84	0.2973
29	83	0.2972
30	82	0.2971
31	81	0.2970
32	87	0.2970
33	82	0.2969
34	85	0.2968
35	68	0.2967
36	82	0.2966
37	82	0.2965
38	82	0.2960
39	82	0.2954
40	87	0.2931

Table 4. *The top of the VFOM* (6) *ranking order for phase sets from SAYTAN for* 2Zn-*insulin with* 1.5 Å *data* (*R* = 5 Å)

Rank number	Mean phase error (°)	VFOM ($\times 10^{6}$)
1	65	0.9130
2	64	0.9120
3	65	0.9105
4	65	0.9099
5	66	0.9097
6	66	0.9094
7	66	0.9094
8	64	0.9074
9	66	0.9037
10	66	0.8966
11	67	0.8944
12	68	0.8934
13	61	0.8907
14	71	0.8895
15	68	0.8890
16	84	0.8723
17	84	0.8717
18	84	0.8624
19	85	0.8610
20	69	0.8584
21	84	0.8574
22	84	0.8508
23	84	0.8485

A further test and discussion

The results with VFOM for the 2 Å aPP phase sets are somewhat better than those obtained by Mukherjee & Woolfson (1993) with modified conventional FOM's. However, the bases of the conventional FOM's are the relationships which exist for small structures, which are known to hold badly for larger structures. In this context aPP may well be near the limit at which they have any useful validity. On the other hand VFOM is based on the properties of the maps of protein structures, that they are divided into distinct regions with distinct properties, and it is known that these properties are valid and useful even for very large structures.

To check on this we applied VFOM to phases derived for 727 large *E*'s by the application of *SAYTAN* to 1.5 Å data for 2Zn-insulin (Mukherjee & Woolfson, 1994). Of the 1000 sets of phases which produced by *SAYTAN* 15 had MPE less than 70° and as will be seen from Table 4 these are comfortably selected by VFOM. Conventional figures of merit were quite useless for this structure so here we have a situation where the statistical relationships encapsulated in *SAYTAN* are able to generate potentially useful phase sets but tests based on similar principles cannot distinguish the good sets from the bad ones. Trials with *SAYTAN* for structures of similar size to 2Zn-insulin, but not containing heavier atoms, have given larger phase errors of order 72°. We suspect that it is the presence of the Zn atoms which gives the comparatively favourable *SAYTAN* outcome for 2Zn-insulin.

We have previously mentioned the log-likelihood criterion based on maximum-entropy extrapolation (Gilmore, Henderson & Bricogne, 1991), which has also been applied to the 0.98 Å resolution phase sets from *SAYTAN*. It is interesting to note that in this procedure only the 117 phases corresponding to reflections of less than 2 Å resolution were kept fixed while the phases of the others were subjected to improvement by maximization of entropy. It seems that a better comparison with the present work could be made if VFOM was applied to phase sets which had previously been subjected to phase refinement but this has not been done. Another point of comparison between the log-likelihood criterion and VFOM involves the contrast between the FOM values for good phase sets and others. The log-likelihood criterion gives extreme differences so that good phase sets are clearly distinguishable whereas it will be clear from Tables 3 and 4 that the differences of VFOM for good and bad sets are small. This may be a cause for concern although the tests reported here were valid ones with observed data and for two structures with quite different characteristics.

It is likely that the maximum-entropy-based FOM would also work at lower resolutions since entropy maximization is a process which has been applied successfully with low-resolution data. Its main drawback is that it is very demanding on computer resources with a single cycle of refinement requiring 14 FFT's compared with five FFT's for VFOM. The CPU time for VFOM evaluation for aPP was under 1 min per set or about 15 h for the 1000 phase sets on a HP730 workstation. While this is quite time consuming it is worthwhile if it leads to the solution of a major structure.

Given the availability of an FOM which is valid for protein phase sets the next consideration is whether information-containing phase sets can actually be obtained on which to apply it. Our experience is that phase sets with MPE's in the range 64–72° can be obtained for proteins with up to about 200 amino acids in the asymmetric unit. If an FOM is available which can recognise these, the the boundary of the usefulness of direct methods will have been moved from what we can recognize to what we can actually produce.

We wish to express our gratitude to the Royal Society which has supported our collaborative work and to the Science and Engineering Research Council for the computer resources without which we could do little. We also thank Dr Laila Refaat and Dr Cecil Tate for their useful comments and assistance.

References

Bhat, T. N. & Blow, D. M. (1982). *Acta Cryst.* **A38**, 21–29.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* **D49**, 148–157.

Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1988). *Acta Cryst.* **A44**, 353–357.

Gilmore, C. J, Henderson, A. N. & Bricogne, G. (1991). *Acta Cryst.* **A47**, 842–846.

Glover, I., Haneef, I., Pitts, J., Wood, S., Moss, T., Tickle, I. & Blundell, T. (1983). *Biopolymers*, **22**, 293–304.

Jones, E. Y., Walker, N. P. C. & Stuart, D. I. (1991). *Acta Cryst.* **A47**, 753–770.

Leslie, A. G. W. (1987). *Acta Cryst.* **A43**, 134–136.

Mukherjee, M. & Woolfson, M. M. (1993). *Acta Cryst.* **D49**, 9–12.

Mukherjee, M. & Woolfson, M. M. (1994). *Acta Cryst.* Submitted.

Reynolds, R. A., Remington, S. J., Weaver, L. H., Fisher, R. G., Anderson, W. F., Ammon, H. L. & Matthews, B. W. (1985). *Acta Cryst.* **B41**, 139–147.

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

Zhang, K. Y. J. & Main, P. (1990*a*). *Acta Cryst.* **A46**, 41–46.

Zhang, K. Y. J. & Main, P. (1990*b*). *Acta Cryst.* **A46**, 377–381.