Acta Cryst. (1998). A**54**, 44–48

Reinvestigation of the Use of Patterson Maps to Extrapolate Data to Higher Resolution

DAVID A. LANGS

Hauptman-Woodward Medical Research Institute, Inc., 73 High Street, Buffalo, NY 14203, USA.
E-mail: langs@hwi.buffalo.edu

(Received 25 February 1997; accepted 22 July 1997)

Submitted on the occasion of Herbert Hauptman's 80th birthday

Abstract

Many years ago, Karle & Hauptman proposed that the Patterson function could be used for data extrapolation beyond the observed range of the actual measured data. Few people have subsequently attempted to exploit this interesting idea, which might suggest possible limitations of this method, even in structural applications of modest complexity. This appears not to be the case, however, but the original ideas for implementing the extrapolation can be significantly improved. New calculation protocols indicate that Patterson maps may be used to extend observed data sets from 1.0 to ~0.5 Å resolution with reasonably good precision. Correlation coefficients between the extrapolated $F(hkl)$'s and their structure-computed expected values typically range between 0.40 and 0.70 across the unobserved range, even for structures containing as many as 600 non-H light atoms in the asymmetric unit. The method is equally good at extrapolating F values for small zones of data that may not have been recorded within the observed resolution range of the diffraction experiment. Furthermore, triplet phase invariants that incorporate one or two extrapolated terms are nearly as reliable as those formed using only the observed data.

1. Introduction

Direct phasing methods have recently experienced a substantial improvement with regard to the size of the structures that can successfully tackled. The *Shake-and-Bake* algorithm (DeTitta, Weeks, Thuman, Miller & Hauptman, 1994; Weeks, DeTitta, Hauptman, Thuman & Miller, 1994) has demonstrated the power of combining reciprocal-space phase refinement with real-space Fourier structure validation in the solution process. This new procedure has worked remarkably well for structures containing as many as 624 non-H atoms (Smith, Blessing, Ealick, Fontecilla-Camps, Hauptman, Housset, Langs & Miller, 1996) provided that native diffraction data are measurable to at least 1.2 Å resolution. Efforts to extend the use of these methods to diffraction data sets of lower resolution are an important consideration with regard to potential applications involving macromolecular structures that do not diffract to atomic resolution. To address this issue, this paper has re-examined procedures that will allow a measured set of diffraction data to be extended to higher resolution with reasonable precision.

2. Background

Karle & Hauptman (1964) proposed an iterative procedure more than 30 years ago, whereby a sharpened origin-removed Patterson function could be utilized to extrapolate data beyond the resolution limit of the observed data. An $|E(hkl)|^2 - 1$ Patterson was computed and modified by zeroing all negative grid-point intensities and expected void regions of the map and then the map was back-transformed to obtain new estimates of the $|E(hkl)|^2 - 1$ coefficients, both for the initial observed data and for lattice points beyond the observed data range. Before computing the next iterative map, one ensured that no $|E(hkl)|^2 - 1$ had values less than -1.0, and then rescaled the extrapolated data in incremental shells of $\sin\theta/\lambda$ to ensure that the shell average $|E(hkl)|^2$ values were approximately 1.0. Values of $|E(hkl)|^2 - 1$ at space-group-extinct positions were presumably omitted, rather than included with default values of -1.0 as would be done for an observed $|E(hkl)|$ of zero. In practice, it was found to be expedient to gradually increase the range of extrapolation by only 10 to 20% after each refinement cycle and the procedure was reported to converge to a stable solution after a limited number of cycles.

The original tests were performed using the $P2_12_12_1$ Cu $K\alpha$ data set for arginine dihydrate; the number of terms was expanded from 1406 measured data to a total of 2688 terms after six refinement cycles. The final $|E(hkl)|^2 - 1$ Patterson map showed remarkably better peak resolution but the actual precision of the extrapolated E values was not objectively quantified. A subsequent paper (Karle & Karle, 1964) indicated that these extrapolated E values were useful for improving estimates of the three-phase structure invariants by means of the $B3.0$ formula (Karle & Hauptman, 1958) but no details of this analysis were given.

One paper (Seeman, Rosenberg, Suddath, Kim & Rich, 1976) reports using this method to obtain better $E(hkl)$ estimates within the observed set of measured data (*i.e.* without extrapolation). The authors were unable to

phase sodium adenylyl-3′,5′-uridine hexahydrate by Fourier methods that recycled the phosphorus site using their original set of scaled E values but were able to obtain a successful convergence after the E values were refined using the Patterson-map procedure described above. Apart from the above-mentioned references, no subsequent work appears to have been pursued to more exactly quantify just how well these procedures may self-validate a measured set of data or extrapolate values beyond the resolution limit of the diffraction experiment.

3. Analysis

A number of critical points were re-examined with regard to assessing the usefulness of Patterson maps in data self-validation and extrapolation. First, does sharpening the $|F(hkl)|^2$ coefficients actually improve the results? Second, how do various Patterson density-modification schemes compare and are there biases imparted to the extrapolated Patterson coefficients using certain schemes? And third, how useful are the extrapolated amplitude data? Can they be used actively in direct-methods phasing or does their limited accuracy only warrant that they be used in a more passive manner? And, lastly, at what limiting initial resolution might these methods prove ineffective?

Our initial survey of these computational variables was performed on a panel of small-molecule Cu $K\alpha$ resolution structures containing 30 to ~100 non-H light atoms. The various refinement strategies were evaluated by comparing the computed correlation coefficients (CC's) between the extrapolated $F(hkl)$ values and either (*a*) their observed measured values or (*b*) their unmeasured but 'known' true values computed from the refined structural parameters.

CC's will have a distinct advantage over conventional R factors in quantifying this agreement. The average values of the F's in each $\sin\theta/\lambda$ range of reciprocal space will differ for observed/known *versus* extrapolated $F(hkl)$ values, depending on the number of iterative cycles of refinement that are performed. In order to get the best R-factor agreement, the average extrapolated F's have to be rescaled to their average known values before the residual is computed, whereas CC's will have the same values regardless of the relative scales between quantities that are compared. A CC of 1.0 would indicate a perfect agreement between the compared values of two sets of data, a zero value would indicate that this agreement is no better than random, negative values would indicate the agreement is worse than random, *i.e.* large F's computing small and small F's computing large.

Patterson coefficients exhibiting five progressive degrees of resolution-dependent sharpening were tested: (I) $|F(hkl)|^2$; (II) $[|F(hkl)|/f(\text{carbon})]^2$; (III) $[|F(hkl)|\exp(B_{\text{eff}}s^2)/f(\text{carbon})]^2$; (IV) $|E(hkl)|^2$; and (V) $|E(hkl)|^2 - 1$. Here, $f(\text{carbon})$ is the atomic scattering factor for carbon or some appropriate weighted average of atomic types for the structure. B_{eff} is an effective isotropic temperature factor for resolution-dependent sharpening, which can be either higher or lower than the average scaled isotropic temperature factor of the data, and s is the value of $\sin\theta/\lambda$ for the particular reflection $F(hkl)$. E values were obtained using the anisothermal scaling features (Blessing & Langs, 1988) built into the *DREADD* set of data-reduction programs developed by Blessing (1989).

The most obvious judgment prior to modifying Patterson maps was first to assess whether the zero-density threshold proposed by Karle & Hauptman (1964) was optimal for discriminating those regions of the map that contained relatively few important Patterson vectors relative to the rest of the map. In addition, might there be better strategies for density modification that did not rely on a fixed threshold criterion? To ensure adequate 'peak' resolution, Patterson maps were computed with a minimal grid-interval size that was at least four times the maximum value of the reflection indices in each of the three dimensions.

4. Results

Resolution-dependent enhancement of the $F(hkl)$ amplitudes markedly improved the accuracy of data extrapolation from the Patterson function. Two undesirable characteristics noted for unsharpened $|F(hkl)|^2$ map refinements could be minimized as the data were progressively sharpened. For the smaller test structures having ~100 or fewer non-H atoms: (*a*) the extrapolated $|F(hkl)|^2$ for unobserved data quickly exceeded their known true target values in as few as three or four cycles of refinement; and (*b*) the most negative features of the initial observed Patterson map tended to persist as refinement proceeded and the contribution of the extrapolated 'unobserved' amplitudes to the synthesis increased. When $[|F(hkl)|/f(\text{carbon})]^2$ coefficients were used: (*a*) the extrapolated magnitudes of the unobserved data grew more slowly and approached their true values in about five to ten cycles; and (*b*) the most negative features in the initial Patterson map tended to become less negative by about 30% when convergence was achieved.

For the third set of tests, protocol (III), the data were further sharpened using an effective isotropic B value. Optimal results were obtained when B_{eff} was set approximately equal to B_{min}, the lower limit of the isotropic B values of the structure. The CC's between extrapolated unobserved data and their known true values were significantly higher as compared to results obtained by protocols (I) and (II) or using B_{eff} values that were as little as ± 1.0 units from B_{min}. For most of the data sets examined, B_{min} was 2.0 to 5.0 units less than $\langle B_{\text{iso}}\rangle$, the average scaled isotropic B value estimated from the data. For the refinements with B_{eff} equal to B_{min}, the $\sin\theta/\lambda$ shell-averaged values of the extrapolated amplitudes of

the unobserved data approached, but usually did not exceed, their true values after 20 or more iterative cycles. At the end of the refinement, the most negative features of the Patterson map (prior to density modification) were usually 50% or less of their values from the initial map computed with only observed data.

Results obtained with protocol (IV), which employed anisotropically scaled $|E|^2$ values were only slightly better or worse than those obtained using protocol (III) with $B_{\rm eff}$ equal to $\langle B_{\rm iso}\rangle$, and significantly worse than those obtained setting $B_{\rm eff}$ equal to $B_{\rm min}$. Results from protocol (V), which used the $|E(hkl)|^2 - 1$ Patterson, were also not quite as good as those obtained by protocol (III) with $B_{\rm eff}$ equal to $B_{\rm min}$.

It also became clear that it was more advantageous to extrapolate values for all data from the beginning rather than to increase the resolution gradually each cycle until about twice the number of data had been accessed. It will be shown that it is often possible to extend the useful range of resolution from twofold to about fivefold the number of data by refining the full set of accessible extrapolated values in each refinement cycle.

Apart from sharpening the coefficients of the Patterson maps, it was quickly determined that better extrapolation results could be obtained if, early in the recycling procedure, the critical threshold was raised significantly above the zero level of the map. As in the case of the earlier studies, maps were computed excluding the $F(000)$ term for this purpose. As a rough rule of thumb, if the most negative feature in the initial observed Patterson map was $-N$ units, it was found effective to zero all the features that were less than a positive threshold of $+2N$ units in the first refinement cycle. Over the next M cycles of refinement, this threshold was progressively lowered by $-3N/M$ units until it reached the value of $-N$ units observed in the original map. More elaborate density-modification schemes were not thoroughly investigated at this time.

This sliding threshold procedure had two major advantages as compared to using a fixed zero threshold in each cycle. First, the percentage of negative extrapolated data rapidly approached zero by the last cycle; this was especially so if one did not allow the observed data terms to refine in each cycle. Second, the accuracy of the extrapolated refined values of the unobserved data was markedly improved as judged by CC's computed by our test examples.

The extrapolation results provided by two large test structures is reported: (i) gramicidin A, $N = 317$ non-H light atoms (Langs, 1988); and (ii) scorpion toxin-II (*Androctonus australis* Hector), $N = 624$ non-H atoms (Fontecilla-Camps, Hebersetzer-Rochat & Rochat, 1988). Both structures are orthorhombic $P2_12_12_1$ and diffract to 0.86 and 0.96 Å resolution, respectively. The gramicidin data set contained 21 454 independent reflections and is essentially complete, lacking only the lowest-order $F(110)$ reflection. The toxin II data set contained 31 001 reflections and is about 90% complete, lacking an inaccessible cone of data coincident with the c axis of the crystal.

Table 1. *Protocol (III) gramicidin extrapolation results are compared to those obtained by the older $|E^2|-1$ procedure*

The data are arranged in approximate equal population shells based on the increasing value of $\sin\theta/\lambda$ (Å^{-1}). The correlation coefficient (CC) and number of data (♯) in each shell are given as well as the cumulative number of data (SUM♯) that are accessible to that limit of resolution.

	Protocol (III)			Old $\|E^2\|-1$ method		
$\sin\theta/\lambda$	♯	CC	SUM♯	♯	CC	SUM♯
0.237	3779	1.000	3779	3711	0.900	3711
0.364	3624	1.000	7403	3559	0.916	7270
0.433	3586	1.000	10989	3538	0.966	10808
0.485	3550	1.000	14539	3529	0.955	14337
0.527	3562	1.000	18101	3555	0.964	17892
0.563	3353	1.000	21454	3347	0.970	21239†
0.596	3531	0.388	3719	3409	0.406	3409
0.625	3520	0.417	7239	3434	0.416	6843
0.652	3498	0.425	10737	3444	0.364	10287
0.677	3539	0.474	14276	3509	0.366	13796
0.700	3482	0.478	17758	3455	0.370	17251
0.721	3513	0.556	21271	3484	0.329	20735
0.741	3500	0.616	24771	3472	0.312	24207
0.761	3505	0.591	28276	3494	0.335	27701
0.779	3485	0.580	31761	3472	0.322	31173
0.796	3464	0.567	35225	3446	0.280	34619
0.813	3543	0.562	38768	3537	0.235	38156
0.829	3461	0.601	42229	3450	0.304	41606
0.845	3482	0.625	45711	3477	0.305	45083
0.860	3451	0.607	49162	3450	0.266	48533
0.874	3514	0.657	52676	3513	0.233	52046
0.888	3491	0.667	56167	3490	0.202	55536
0.902	3478	0.667	59645	3479	0.197	59015
0.915	3493	0.698	63138	3491	0.183	62506
0.928	3453	0.713	66591	3454	0.206	65960
0.940	3503	0.674	70094	3509	0.144	69469
0.952	3456	0.691	73550	3457	0.125	72926
0.964	3451	0.662	77001	3460	0.100	76386
0.972	904	0.692	77905	904	−0.031	77290

† Limit of observed data.

Both data sets were refined by protocol (III) with $B_{\rm min}$ equal to 4.0 Å^2 using the sliding density-modification threshold described above. Patterson maps for both structures were computed on a 128 × 128 × 128 grid and the observed data were extended to a maximum resolution of 0.5 Å. The gramicidin data required 20 cycles of refinement for optimal convergence, the larger toxin II structure required 40 cycles. Parallel calculations were performed for the gramicidin data using the $|E|^2 - 1$ Patterson method outlined by Karle & Hauptman (1964). CC comparisons for the extrapolated data of the two gramicidin refinements are shown as a function of increasing $\sin\theta/\lambda$ in Table 1. The toxin-II extrapolation results are given in Table 2.

An important question to be answered is whether these extrapolated E values form reliable triplet invariants that are good enough to be actively used in a direct-methods

Table 2. *Toxin-II extrapolation results after 40 refinement cycles*

Columns are labeled similar to those in Table 1.

Observed agreement				Extrapolated agreement			
$\sin\theta/\lambda$	♯	CC	SUM♯	$\sin\theta/\lambda$	♯	CC	SUM♯
0.214	5804	1.000	5804	0.158	211	0.698†	211
0.325	5597	1.000	11401	0.328	218	0.211	429
0.387	5452	1.000	16853	0.388	307	0.054	736
0.433	5249	1.000	22102	0.435	445	0.146	1181
0.470	4800	1.000	26902	0.472	917	0.175	2098
0.502	4099	1.000	31001	0.507	1575	0.104	3673

† Agreement for the lowest-resolution shell of extrapolated data from the missing data of unmeasured reflections is especially good.

Table 3. *Breakdown of gramicidin triples containing 0, 1, 2 or 3 extrapolated E magnitudes*

The analysis is given for three progressively larger groups of triples for which $A \geq 0.8$, 0.5 and 0.4, respectively. $A = 2|E_h E_k E_l|/N^{1/2}$, where $h + k + l = 0$ and N is approximately equal to the number of equivalent non-H atoms in the unit cell. The number of triples for each group based on the number of extrapolated E values (♯$X - E$), the average A value $\langle A\rangle$ and the average theoretically expected and average actual true cosine invariant values for each group, $\langle\varepsilon(\cos\Phi)\rangle$ and $\langle\cos\Phi_{tr}\rangle$, are listed. The triplet estimates are particularly good when the ratio $\langle\cos\Phi_{tr}\rangle/\langle\varepsilon(\cos)\rangle$ is equal to or greater than 1.0. The estimates may be deemed seriously in error when this ratio is significantly less than 0.5 or even negative as is indicated by the daggers (†) in the rightmost column of the table.

	♯$X - E$	♯triples	$\langle A\rangle$	$\langle\varepsilon(\cos\Phi)\rangle$	$\langle\cos\Phi_{tr}\rangle$	Ratio
$A \geq 0.8$	0	2526	1.03	0.472	0.475	1.01
	1	1143	1.08	0.491	0.444	0.90
	2	1582	0.997	0.455	0.311	0.68
	3	23	0.893	0.409	−0.402	−0.98†
$A \geq 0.5$	0	17805	0.658	0.330	0.320	0.97
	1	9413	0.648	0.327	0.260	0.80
	2	11866	0.651	0.324	0.196	0.60
	3	620	0.578	0.282	0.071	0.25†
$A \geq 0.4$	0	40546	0.537	0.278	0.264	0.95
	1	26146	0.515	0.266	0.197	0.74
	2	27790	0.530	0.272	0.157	0.58
	3	2142	0.479	0.238	0.073	0.31†

† Ratio is significantly less than 0.5.

phase determination to improve our chances of obtaining solutions. Table 3 lists the average values of the gramicidin cosine phase invariants, $\langle\cos_{tr}\rangle$, for triples that contain one, two and three extrapolated E values and compares them to their average $I_1(A)/I_0(A)$ expected values $\langle(\cos)\rangle$.

5. Discussion of results

Table 1 shows that the CC comparison for extrapolated data is significantly better using protocol (III) as compared to the older $|E^2| - 1$ method. Values for the new procedure vary from 0.39 to 0.71 in comparison to 0.41 to −0.03 from the older method for the unobserved extrapolated range from 0.86 to 0.5 Å resolution. It would appear to be logical that this agreement should be best for those extrapolated data that are closest to the observed sphere of measured data, as is indicated by the $|E^2| - 1$ analysis. It is quite surprising that the protocol (III) CC's actually improved quite significantly at higher resolution, and had the four highest CC values (0.698, 0.713, 0.674, 0.691) in the four contiguous resolution shells in the range 0.525–0.546 Å (0.915–0.952 Å^{-1}). Perhaps this is due to the average lower temperature of the main-chain atoms of the helical backbone of this structure, which is perhaps more dominant in the scattering at higher resolution, but it does not explain why the $|E^2| - 1$ method cannot take advantage of this fact. It may be noted from the toxin-II test example in Table 2, however, that extrapolation to higher resolution may be quite difficult if a significant percentage of the low-resolution terms in the observed Patterson map are missing. Although the lowest-resolution shell of the missing data appears to be reliably estimated with a CC of 0.698, *i.e.* the 211 terms indicated by the dagger in Table 2, CC's beyond this resolution limit seldom exceed 0.20 as long as these potentially large magnitude terms are omitted from our initial Patterson map.

The original gramicidin structure determination used 1500 phases and 17 500 triples with $A \geq 0.5$. The tangent-formula solution gave 1320 phases with an absolute mean phase error of 39°. The E map revealed 105 true atom sites in the top 150 peaks. If more than 1500 phases or triples having A values less than 0.5 are used, the solution rapidly degrades and becomes unidentifiable. This instability is demonstrated by inputting the known phases of the structure and performing a number of cycles of tangent-formula phase refinement until convergence is noted.

After data extrapolation, the top 3000 E values were used to generate 96 624 triples for which A was equal to or greater than 0.4. These 3000 E values consisted of 1519 observed and 1481 extrapolated, previously unobserved, data. The triples that contained as many as three extrapolated E values were considered to be too unreliable, as is indicated by the bottom-most dagger in Table 3, so these 2142 triples were deleted from the list prior to phasing. The solution now gave 2599 phases with a mean absolute phase error of 35°. If only those 1477 terms among the original top 1500 E values are examined, the mean absolute phase error is only 28°, as compared to 39° obtained previously. The resulting E map revealed 119 atoms in the top 150 peaks. Thus, it has been demonstrated that a smaller phase error and better map details can result from using additional extrapolated data in the direct-methods process.

Efforts to extend the use of these methods to initial data sets of lower resolution, say 1.5 Å, have not been as successful but will doubtless have an important impact on direct-methods phasing applications of larger macromolecular structures. Although the CC's for 1.5 Å resolution test structures seldom exceeded 0.10 for all but the closest shell of extrapolated values nearest to the

observed data, we are hopeful that more powerful strategies may be developed to further this goal.

We thank Drs Juan Carlos Fontecilla-Camps and Dominique Housset for the use of the 0.96 Å scorpion toxin-II data. Results of the *ab initio* direct-methods determination of this 624-atom structure are reported elsewhere (Smith, Blessing, Ealick, Fontecilla-Camps, Hauptman, Housset, Langs & Miller, 1997). Research support from NIH grant GM-46733 is gratefully acknowledged.

References

Blessing, R. H. (1989). *J. Appl. Cryst.* **22**, 396–397.
Blessing, R. H. & Langs, D. A. (1988). *Acta Cryst.* A**44**, 729–735.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Fontecilla-Camps, J. C., Hebersetzer-Rochat, C. & Rochat, H. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 7443–7447.
Karle, I. L. & Karle, J. (1964). *Acta Cryst.* **17**, 835–841.
Karle, J. & Hauptman, H. (1958). *Acta Cryst.* **11**, 264–269.
Karle, J. & Hauptman, H. (1964). *Acta Cryst.* **17**, 392–396.
Langs, D. A. (1988). *Science*, **241**, 188–191.
Seeman, N. C., Rosenberg, J. M., Suddath, F. L., Kim, J. J. P. & Rich, A. (1976). *J. Mol. Biol.* **104**, 109–144.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1996). *Acta Cryst.* A**52**, C64–C65.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Acta Cryst.* D**53**, 551–557.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.