Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 24 February 2011
Accepted 11 April 2011

Submitted on the occasion of Herbert Hauptman's 94th birthday

Relaxation of the resolution requirements for direct-methods phasing

David A. Langs* and Herbert A. Hauptman

Department of Structural Biology, Hauptman–Woodward Medical Research Institute, 701 Ellicott Street, Buffalo, NY 14203, USA. Correspondence e-mail: langs@hwi.buffalo.edu

Shake-and-bake phasing methods have permitted the *ab initio* solution of crystal structures containing more than 1000 independent non-H light atoms (C, N, O). The success of these procedures is critically dependent upon having diffraction data measured to at least 1.2 Å resolution. A new target function $R_2(\varphi_{\mathbf{h}})$ is introduced into the shake-and-bake procedure along with a real difference map strategy whereby this resolution limit can be appreciably lowered toward 1.5 Å. These improvements, when applied to moderately high resolution data, may now allow one the possibility to solve structures that are twice as large as could have been solved previously.

1. Introduction

Whereas the tangent formula (Karle & Hauptman, 1956) had dominated the early history of direct-phasing methods, dual-space shake-and-bake (SnB) procedures (Hauptman, 1988; DeTitta *et al.*, 1994; Weeks *et al.*, 1994) have supplanted the tangent formula in recent years. Although SnB has permitted a five- to tenfold increase in the size and complexity of structures that could be solved by tangent-formula methods, it has not improved upon the minimal resolution of the data required for *ab initio* direct methods to succeed, namely 1.2 Å. Other similar sophisticated direct-methods programs such as *SHELXD* (Usón & Sheldrick, 1999), *ACORN* (Foadi *et al.*, 2000), *SIR/IL MILIONE* (Burla *et al.*, 2005; Caliandro *et al.*, 2009) and *VLD* (Burla *et al.*, 2011) have also wrestled with this issue. This resolution limit can occasionally be relaxed for structures containing heavier atoms such as sulfur and larger, especially if the locations of these heavy atoms can be previously determined by other methods. The efforts to further develop direct-phasing methods for native light-atom crystal structures have thus far characteristically offered only marginal improvements with regard to relaxing the 1.2 Å resolution limit. The *ARCIMBOLDO* program (Rodríguez *et al.*, 2009) might be considered to be a quasi *ab inito* direct-phasing method. *ARCIMBOLDO* can solve protein structures at resolutions as low as 2 Å provided that small generic α-helices, of sufficient scattering power, can be correctly positioned in the cell to initiate the phasing process.

2. Background

The SnB procedure operates in both real and reciprocal space. It begins with a randomly generated N-atom starting model where N is approximately equal to the number of independent non-H atoms in the asymmetric portion of the unit cell. Given these atomic coordinates, phases are computed for the $\sim 10N$ largest $|E|$ values. Approximately $100N$ phase invariants ($\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}}$) are generated for the basis of the reciprocal-space refinement. Each phase is refined to minimize $R_1(\varphi_{\mathbf{h}})$, a residual between its collection of computed cosine invariants, $\cos(\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}})$, and τ, their expected values,

$$R_1(\varphi_{\mathbf{h}}) = \sum_{\mathbf{k}} A_{\mathbf{h},\mathbf{k}} \left[\cos(\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}}) - \tau\right]^2 / \sum_{\mathbf{k}} A_{\mathbf{h},\mathbf{k}}. \quad (1)$$

Here $A_{\mathbf{h},\mathbf{k}} = 2\sigma_3 |E_{\mathbf{h}} E_{\mathbf{k}} E_{\mathbf{k}-\mathbf{h}}| / \sigma_2^{3/2}$, where $\sigma_n = \sum_{j=1,N_{\rm cell}} f_j^n$ is summed over $N_{\rm cell}$, the total contents of the primitive unit cell, represented by the atomic form factors f_j raised to the appropriate power. The most common refinement scheme, often referred to as quadrant permutation, first calculates $R_1(\varphi_{\mathbf{h}})$ given the initial atom-based values of the phases, then recalculates this function three other times after incrementing the value of $\varphi_{\mathbf{h}}$ by ± 90 and 180°, and then accepts the value of $\varphi_{\mathbf{h}}$ producing the lowest $R_1(\varphi_{\mathbf{h}})$ value. The phases are sequentially refined for a small number of cycles prior to computing an E map and selecting the N largest peaks for the next overall cycle of refinement. After 100 or so overall refinement cycles have been completed, $R(\varphi)$ is computed for the entire phase set as $R_{\rm min}$. After a sufficient number of different random-atom refinement trials have been completed, one can inspect a histogram of the $R_{\rm min}$ values to identify solutions.

Earlier studies investigated the effect of different parameter-shift protocols for sampling $R_1(\varphi_{\mathbf{h}})$ and determining its minimum value. Higher success rates were noted if zonal restricted phases were allowed to vary as unrestricted values in the earlier cycles of refinement. Different target functions were also investigated with regard to increasing the success rates and overall efficiency of the phase-determination process. These have included the sine-enhanced (Xu *et al.*, 2002) exponential (Hauptman *et al.*, 1999) and statistical (Xu & Hauptman, 2004) versions of the SnB program. These calculations were performed on a battery of representative

light-atom test structures of various sizes in different space groups, all at moderately high resolution, usually less than 1.0 Å. Similar trials were also run using ΔE derivative data for macromolecular heavy-atom substructures, usually between 3 and 4 Å resolution, but these shall not concern us here (Xu *et al.*, 2002; Xu & Hauptman, 2006).

In most instances the original $R_1(\varphi_{\mathbf{h}})$ target function significantly outperformed most of the others, but on occasion a particular chosen variant could do better under certain conditions. In all fairness, however, these studies were not extended to lower-resolution light-atom data sets at which point these methods begin to fail. Recent work now suggests that the effectiveness of these alternative target functions should be reinvestigated with regard to applications to lower-resolution data.

3. Sine-squared target function

Attention here is to be called to a particular target function, originally proposed by Hauptman more than 15 years ago, which in previous tests at moderately high resolution was much less effective than the $R_1(\varphi_{\mathbf{h}})$ function, which typically had a success rate which was four or more times higher. The new function adds a $\sin^2(\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}})$ penalty component to the $R_1(\varphi_{\mathbf{h}})$ minimization that tends to drive $\cos(\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}})$ toward larger values to accelerate the refinement process. To wit

$$R_2(\varphi_{\mathbf{h}}) = \sum_{\mathbf{k}} A_{\mathbf{h},\mathbf{k}} \Big\{ \big[\cos(\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}}) - \tau\big]^2 + \sin^2(\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}}) \Big\} / \sum_{\mathbf{k}} A_{\mathbf{h},\mathbf{k}}. \quad (2)$$

It should be noted that this equation can be rewritten as

$$R_2(\varphi_{\mathbf{h}}) = \sum_{\mathbf{k}} \big[A_{\mathbf{h},\mathbf{k}}(1 + \tau^2) - \alpha_{\mathbf{h}} \cos(\varphi_{\mathbf{h}} - \chi_{\mathbf{h}})\big] / \sum_{\mathbf{k}} A_{\mathbf{h},\mathbf{k}}, \quad (3)$$

where the values $\alpha_{\mathbf{h}} = (A^2 + B^2)^{1/2}$, $A = \sum_{\mathbf{k}} \tau A_{\mathbf{h},\mathbf{k}} \cos(\varphi_{\mathbf{h}-\mathbf{k}} + \varphi_{\mathbf{k}})$, $B = \sum_{\mathbf{k}} \tau A_{\mathbf{h},\mathbf{k}} \sin(\varphi_{\mathbf{h}-\mathbf{k}} + \varphi_{\mathbf{k}})$ and $\chi_{\mathbf{h}} = \tan^{-1}(B/A)$. Unlike $R_1(\varphi_{\mathbf{h}})$, it is clear to see that $R_2(\varphi_{\mathbf{h}})$ is simply a τ-weighted tangent function which is bimodal, and has a minimum when $\varphi_{\mathbf{h}} = \chi_{\mathbf{h}}$ and a maximum when $\varphi_{\mathbf{h}} = \chi_{\mathbf{h}} + 180^\circ$. It follows that the best phase-shift estimates occur when $\alpha_{\mathbf{h}}$ is large, and the poorest when $\alpha_{\mathbf{h}}$ approaches zero, in which instance all values of $R_2(\varphi_{\mathbf{h}})$ are essentially the same.

4. Initial test calculations

Previous SnB calculations at moderately high resolution have indicated that the higher-symmetry space groups are a bit more difficult to determine than similar-sized structures in lower-symmetry groups such as $P1$ and $P2_1$. For this purpose four $P2_12_12_1$ structures, each of approximately 100 non-H all light (C, N, O) atoms, were selected (ILED, FILE4, FILE5, TERN). The stability of the solution values of the phases to SnB refinement was tested for both the R_1 and R_2 target functions by performing a sufficient number of refinement cycles to ensure that a stable minimum in R had been obtained. These calculations were performed for each diffraction data set truncated at 1.2 $\rightarrow$ 1.5 Å, in 0.05 Å increments. In all cases $\varphi_{\mathbf{h}}$ was permuted by 0, $\pm$90 and 180° as previously described, with three consecutive passes through the list of phases per SnB refinement cycle. When each refinement converged, the average phase error $\delta\varphi$ between the refined phases and their true values was computed, as is recorded in Table 1. Two values for $\delta\varphi$ are reported: the first, $\delta\varphi_1$, is determined from the values of the phases computed from the N peaks selected from the E map at the last refinement cycle; the second, $\delta\varphi_2$, reports the error from the values of the phases obtained at the end $R(\varphi)$ minimization. $\delta\varphi_1$ is usually 5 to 10° less than $\delta\varphi_2$. As such, $\delta\varphi_2$ may be considered as the closest that any particular target function will converge towards the true values of the phases. The normal SnB default, however, is to list the x, y, z coordinates of the N largest peak positions obtained from the final E map rather than the values of the phases obtained at any particular real-/reciprocal-space refinement stage.

Table 1
SnB phase errors for the four structures ILED (N = 84), FILE4 (N = 88), FILE5 (N = 127) and TERN (N = 110).

The solution sets of phases were refined for 100 to 200 SnB cycles for both the R_1 and R_2 target functions at seven different resolutions from 1.2 to 1.5 Å. An asterisk (*) indicates that the phases degraded toward random values with errors exceeding 80°. Trial sets, for which the phase error is less than 45°, can be readily identified by a lower $R_{\rm min}$ value as compared with random non-solution sets. Once the phase error exceeds ~50°, $R_{\rm min}$ loses it ability to clearly identify solutions from non-solutions, but they still usually place within the top 1 or 2% of the sorted $R_{\rm min}$ list. Two values for the phase error ($\delta\varphi_1$; $\delta\varphi_2$) are reported, as explained in the text.

RES (Å)	R_n	ILED†	FILE4‡	FILE5§	TERN¶
1.2	R_1	27; 38	34; 41	38; 44	*
	R_2	19; 29	24; 31	29; 34	44; 48
1.25	R_1	38; 42	*	*	*
	R_2	29; 36	32; 40	32; 34	58; 60
1.30	R_1	*	*	*	*
	R_2	34; 37	34; 38	40; 45	*
1.35	R_2	35; 39	37; 42	46; 49	*
1.40	R_2	38; 47	46; 50	56; 61	*
1.45	R_2	46; 49	*	*	*
1.50	R_2	49; 53	*	*	*

† Pletnev *et al.* (1980). ‡ Pletnev *et al.* (1991). § Pletnev *et al.* (1992). ¶ Miller *et al.* (1993).

Both R_1 and R_2 were next used to solve the ILED structure at 1.2 Å. 2000 random SnB trials were each iterated for 200 refinement cycles using 800 phases, 8000 triples and 80 picked peaks. The R_1 target function produced 68 solutions for the 1.2 Å data as compared with only 18 solutions for the R_2 function, which was noted to be similar to the 4:1 solution ratios usually seen for higher-resolution data sets when these two functions were tested. But it was soon realized that there was a serious flaw in the manner in which R_2 was used to refine the zonal restricted phases of the structure. Although quadrant permutation with the R_1 function allows zonal phases the possibility to occasionally refine away from their restricted values to escape false $R(\varphi)$ minima to significantly improve the success rate, it was belatedly recognized that the same does not hold with regard to the R_2 function. Rather, when zonal

phases are permuted either $\pm 90^\circ$ from their permissible values, the computed $\sin^2(\varphi_{\mathbf{h}} + \varphi_{-\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}})$ component will tend to maximize the value of R_2, thus strongly preventing any temporary refinement towards those values. But when one uses R_2 to refine all the general phases and R_1 to refine only the small subset of zonal phases to allow them to temporarily drift from their restricted plane, the number of solutions produced by this modified procedure is seen to dramatically increase from 18 to 76, thus slightly bettering the 68 successes noted for the R_1 trials cited above. In a previous study the tangent formula was shown to be as effective as the R_1 target function in SnB applications to small- or medium-size structures having high-resolution data (Chang *et al.*, 1997). But in comparison, the tangent formula only produced 25 solutions in 5000 random SnB trials when applied to the 1.2 Å ILED data, which is to say about seven times less effective than either the R_1 or R_2 functions, while the $\chi_{\mathbf{h}}$ target function proved significantly better with 37 solutions in 5000 sets.

The R_2 refinements were repeated for the ILED data truncated at 1.4 Å using 600 phases, 10 000 triples and 65 peaks, for it was unreasonable to expect that all 84 atoms could be resolved from the E maps at this lower resolution. Whereas now the R_1 target function does not produce any solutions, the modified R_2 scheme produced 11 solutions with $\delta\varphi_1$ ranging between 38 and 50°, the majority of which could be tentatively identified by their lower $R_{\min}$ values. Random-atom SnB trials were also computed for the 1.5 Å ILED data (600 phases, 8000 triples, 65 peaks), and produced nine sets with $\delta\varphi_1$ ranging between 49 and 59°, but $R_{\min}$ was not as selective to identify them as solutions among the 2000 sets.

5. Preliminary concerns

In Table 1 it is observed that the $\delta\varphi_1$ values are generally 5 to 10° smaller than their corresponding $\delta\varphi_2$ values, therefore it is evident that the $R(\varphi)$ refinements of the map-based phase values are actually diverging toward their minimal function stable values. Not so astonishing, perhaps, was the fact that the E maps calculated from these degraded phase values were sufficiently good enough to interpolate reasonably accurate peak positions for a large fraction of the N expected atom sites to essentially recover the phase precision observed in the previous cycle. Two questions logically present themselves. Firstly, how can the convergence of the $R(\varphi)$ functions be improved, and, secondly, might there be a better way to determine more of the correct atomic positions from the map density at moderately lower resolution.

6. Improving the $R_2(\varphi_{\mathbf{h}})$ convergence

The normal SnB quadrant permutation scheme may be very effective in navigating the initial random-atom phases toward their solution values for moderately high resolution data, but a serious problem appears to exist with regard to refinement convergence in applications to lower-resolution data. The refinement results in Table 1 for the ILED 1.2 Å data will be more carefully examined (R_1, $\delta\varphi_1 = 27$, $\delta\varphi_2 = 38^\circ$; R_2, $\delta\varphi_1 = 19$, $\delta\varphi_2 = 29^\circ$). To this end Table 2 presents the average values of $\delta\varphi_1$ and $\delta\varphi_2$ for ten groups of 80 phases sorted in descending order on the magnitude of $\alpha_{\mathbf{h}}$. This display clearly shows that the majority of the phases with $\alpha_{\mathbf{h}}$ exceeding 0.69 are well within 45° of their solution values, and will not be further refined by the quadrant permutation protocol. It is the remaining quarter of the phases that have the lowest $\alpha_{\mathbf{h}}$ values that will be most affected by the refinement process, and, indeed, as can be seen for the last two rows having the lowest $\langle\alpha_{\mathbf{h}}\rangle$, the results are disastrous in that the atom-based values are clearly superior to their parameter-shifted refined values. This may be in part due to the relatively small number of three-phase triples invariants that affect the refinement of these phases. If the 10:1 ratio of triples to phases is increased, say to 15:1, the results are significantly improved as is shown for groups of phases in the lower part of Table 2(*b*). In fact, if one now assigns individual weights for the phases as $\mathrm{wt}_{\mathbf{h}} = \tanh(\alpha_{\mathbf{h}})$ and performs several cycles of a weighted R_2 refinement to determine the best values for $\chi_{\mathbf{h}}$ to replace $\varphi_{\mathbf{h}}$, a significantly lower value for $\delta\varphi_2$ is obtained. Whereas previously from Table 1, $\delta\varphi_1 = 19$, $\delta\varphi_2 = 29^\circ$ for the 1.2 Å data, these values can now be significantly reduced to $\delta\varphi_1 = 15$, $\delta\varphi_2 = 19^\circ$. The effects of this weighted $\chi_{\mathbf{h}}$ refinement can also be illustrated with the 1.5 Å ILED data. If we examine the best phase set ($\delta\varphi_1 = 49^\circ$, $\delta\varphi_2 = 53^\circ$) obtained from the 2000 random-atom trials described above, several cycles of weighted $\chi_{\mathbf{h}}$ refinement are sufficient to reduce the phase error to $\delta\varphi_1 = 42^\circ$, $\delta\varphi_2 = 47^\circ$, while its rank improves from 19th to 2nd in the sorted list of 2000 $R_{\min}$ values.

Table 2
The average values of $\delta\varphi_1$ and $\delta\varphi_2$ for the ILED phase refinements presented as ten groups of phases sorted in descending order on the value of $\alpha_{\mathbf{h}}$.

The analysis in columns (*a*) refers to R_1 refinements and in (*b*) to R_2 refinements (both for 1.2 Å data with 800 phases, 8000 triples, 80 atoms picked from map) while columns (*c*) list similar results for the R_2 refinements of the 1.4 Å ILED data (600 phases, 6400 triples, 65 atoms).

(*a*)			(*b*)			(*c*)		
$\langle\alpha_{\mathbf{h}}\rangle$	$\delta\varphi_1$	$\delta\varphi_2$	$\langle\alpha_{\mathbf{h}}\rangle$	$\delta\varphi_1$	$\delta\varphi_2$	$\langle\alpha_{\mathbf{h}}\rangle$	$\delta\varphi_1$	$\delta\varphi_2$
11.71	11	11	13.02	8	8	7.13	14	12
4.28	14	17	4.51	12	12	2.46	23	23
2.58	18	23	2.74	17	18	1.49	38	42
1.77	19	27	1.82	17	19	0.96	25	35
1.25	25	35	1.32	14	18	0.71	38	46
0.90	28	34	0.97	18	23	0.51	48	46
0.69	33	45	0.70	23	31	0.39	40	57
0.52	40	46	0.50	27	40	0.30	51	65
0.32	45	61	0.33	30	46	0.21	45	64
0.16	42	79	0.16	23	71	0.11	57	80
$\langle\delta\varphi^\circ\rangle$	27	38		19	29		38	47

7. Improved map interpretation

At data resolutions greater than 1.2 Å the interpretation of E maps as discrete resolved atomic peaks begins to degrade. For a large peptide or protein molecule lacking sulfur-containing amino-acid residues, the peptide nitrogen and carbonyl

oxygen atoms are generally well resolved and appear as the largest density features in the E map. But the adjacent carbonyl carbon positions, being 1.2 Å from the oxygen and 1.35 Å from the nitrogen positions, will most often not be resolved as a discrete peak. They will at best appear within positive density at a saddle-point minimum between the larger densities that surround the peptide oxygen and nitrogen sites. For example, only about four of the 12 ILED carbonyl C atoms are usually resolved at 1.2 Å, leading to ∼64 correct and ∼20 incorrect peaks among the 84 largest peaks listed.

One obvious way to remedy this situation could be the use of difference maps, with coefficients $(|E_{\mathbf{h}}| - |E_{\mathbf{h}\mathrm{cal}}|)\exp(i\varphi_{\mathbf{h}\mathrm{cal}})$, where $E_{\mathbf{h}\mathrm{cal}}$ is the quasi-normal E value computed from the resolved peaks of the original E map that one wishes to subtract to reveal the locations of atoms at saddle-point densities which cannot normally be determined by the ellipsoidal interpolation algorithm (Rollett, 1965). In practice, however, these difference maps often exhibit ghost peaks for atoms incompletely subtracted and other spurious features as a consequence of restricting the synthesis to only the largest $E_{\mathbf{h}}$ magnitudes. Efforts to improve the scaling between the observed $|E_{\mathbf{h}}|$ and $|E_{\mathbf{h}\mathrm{cal}}|$ only lead to marginal improvements with regard to locating atoms that were missed in the original density.

If, however, one simply edits the original E map to subtract a crude Gaussian-shaped density for each of the first 84 peaks in the map, and then reinterprets the edited map, the results are significantly improved. A simple correction of the form

$$\rho(r)_{\mathrm{dif}} = \rho(r) - \rho_{\mathrm{peak}}\cos(90\delta r/1.2) \qquad (4)$$

will suffice for all grid point distances δr that are less than 1.2 Å from the interpolated peak height of ρ_{peak}. Note that this difference map does not require an additional Fourier synthesis. In the case of the 1.2 Å ILED solution, if the list of main peaks and difference peaks are combined and then sorted in descending order on their extrapolated peak magnitudes, one now observes 75 correct and nine incorrect peaks as compared with 64 correct and 20 incorrect as noted previously, thus reducing the number of incorrect entries from 20 to nine. For the 1.4 Å phase solution noted in Table 1 which has $\delta\varphi_2 = 47^\circ$, only 35 of the top 84 peaks selected from the map correspond to actual atomic locations. After applying the difference density technique, that number is increased from 35 to 40 atoms from the top 84 peaks in the merged resorted list. In this manner by means of $\chi_{\mathbf{h}}$ refinement and the use of difference maps the phase error can be further reduced, at which point the 1.4 Å ILED solutions can be readily identified by their significantly lower R_{min} values.

8. Applications to larger structures

In a practical sense, structures in lower-symmetry space groups are easier to solve than those in higher-symmetry space groups. This is because the magnitude and significance of the A values, $A \simeq 2|E_{\mathbf{h}}E_{-\mathbf{k}}E_{\mathbf{k}-\mathbf{h}}|/N_{\mathrm{cell}}{}^{1/2}$, is inversely dependent on the square root of the total number of light non-H atoms in the primitive unit cell, $N_{\mathrm{cell}} = N_{\mathrm{sym}}N_{\mathrm{asym}}$, where N_{sym} is the number of equivalent positions for the space group and N_{asym} is the total number of independent atoms in the asymmetric unit. It follows that the A values for a triclinic structure are on average twice as large as for the same structure crystallized in the space group $P2_12_12_1$ where $N_{\mathrm{sym}} = 4$. This correlates fairly well with the database of SnB successes reported on our institutional website (http://www.hwi.buffalo.edu/SnB/SnBSuccesses.htm). Since the largest $P1$ structures have about 1300 atoms as compared with around 620 atoms for several $P2_12_12_1$ studies, one can approximate the limiting number of atoms as $N_{\mathrm{limit}} \simeq 1250/N_{\mathrm{sym}}{}^{1/2}$. Larger structures have been reported elsewhere (*e.g.* Pal *et al.*, 2008; Bunkóczi *et al.*, 2005) but have included determinable sulfur sites or large solvent voids that may have aided the phasing process.

Table 3
SnB stability tests for four large $P2_12_12_1$ structures at moderately high resolution: TOX2, AXAN, TONG and AXES.

Only the smaller TOX2 structure is solvable using the standard SnB R_1 target function; the asterisks in line 5 for the AXAN, TONG and AXES structures indicates they are not. Target function R_2 is compared with respect to the phase error ($\delta\varphi_1$; $\delta\varphi_2$) as described in Table 1. The number of E values, triples and peaks selected from the E map for the SnB process are noted. R_3 represents the $\delta\varphi$ values following $\chi_{\mathbf{h}}$ refinement and weak triples augmentation. No difference density mapping is required at this resolution.

		TOX2†	AXAN‡	TONG§	AXES¶
Number of atoms		624	1000	1100	1600
Number of peaks		500	800	800	1000
Number of φ's		5000	8000	12000	12000
Number of triples		50000	185000	260000	200000
All data	R_1	34; 56	*	*	*
	R_2	28; 45	38; 52	36; 54	35; 56
	R_3	25; 42	31; 50	32; 52	28; 48

† Smith *et al.* (1997). ‡ S. D. Trakhanov, V. Z. Pletnev & A. P. Kuzin (unpublished work). § Tong *et al.* (1996). ¶ Ghosh *et al.* (1999).

Here we shall test the effect of applying our new lower-resolution phasing strategies to moderately high resolution data sets for larger $P2_12_12_1$ structures that are normally insolvable to the SnB process. Data are presented in Table 3 for TOX2 (N = 624, 0.96 Å), AXAN (N = 1000, 0.98 Å), TONG (N = 1100, 0.99 Å) and AXES (N = 1600, 0.90 Å). Only the TOX2 structure had been previously solved by the SnB process; the other three were known to be unstable when SnB was seeded with the solution values of the phases. Note that the original SnB R_1 solution of TOX2 (Smith *et al.*, 1997) reported a mean phase error of 19°, rather than the 34° value noted in the table, but this was only after the original 500 solution peaks were Fourier refined for a number of cycles. The best measure of the effectiveness of the various SnB phase-refinement target functions can only be fairly appraised by δ_1 phase errors computed from atoms that have not been extensively Fourier refined, as is the case during iterative SnB refinement.

The number of phases, triples and map peaks selected for the phase-determination process are indicated in the table. Stability tests were initiated with phases computed from a fragment of the structure, and 50 to 100 cycles of refinement were performed to see if a stable solution could be retained. The values of $\delta\varphi_1$ and $\delta\varphi_2$ are noted, as was for the lower-

resolution applications to the smaller structures. Note that the AXAN, TONG and AXEL structures, which were previously unstable to R_1 refinement, now exhibit stable solution minima for the R_2 target function ($\sim$36, $\sim$54°) which are similar to those previously noted for the R_1 solution of the smaller TOX2 structure (34, 56°). We here note furthermore, after employing the $\chi_{\mathbf{h}}$ refinement and difference-density peak search, the results entered as R_3 in Table 3 are appreciably better, particularly so for TOX2 (22, 42°) and the larger AXES structure (28, 48°). Where previously 600 atoms appeared to be the limit for the standard R_1 SnB target function for a $P2_12_12_1$ structure, it now appears that that limit may be increased to more than 1600 atoms by incorporating the R_2 function and strategies described above.

In closing, it is cautioned that, as larger structures are investigated, the range of all $R_{\rm min}$ values, both solutions and non-solutions, strongly converge toward the minimal random expected value of 0.50 (DeTitta *et al.*, 1994). For example, when 2750 random trials were each run 600 SnB cycles for the AXES structure, all non-solution $R_{\rm min}$ values fell between 0.487 and 0.493, with a σ of $\pm$0.001. Although the solution produced the lowest $R_{\rm min}$ value of 0.487, many non-solutions had $R_{\rm min}$ values that were only infintesimally larger. To remedy this problem, a chemically logical discriminant based on reasonable bond-angle geometries was subsequently defined that could clearly identify the true solution from all the non-solution sets,

$$R_{\rm angle} = N_{\rm bad}/\big(N_{\rm good} + N_{\rm bad}\big). \qquad (5)$$

Here, $N_{\rm good}$ is the total number of reasonable bond angles lying between 95 and 135°, while $N_{\rm bad}$ is the number lying outside this range. In this regard, the AXES structural solution produced an $R_{\rm angle}$ equal to 0.23, while the remaining 2750 non-solutions produced a Gaussian $R_{\rm angle}$ distribution ranging between 0.46 and 0.74, thus clearly separating the solution from all non-solution sets having similar $R_{\rm min}$ values.

9. Summary

Solutions for the R_2 target function have been shown to have a greater phase stability and convergence power than for the older R_1 function in applications to lower-resolution data sets and larger structures recorded to moderately high resolution. Crude quadrant permutation is 5 to 10 times more effective than directly seeking the minimum of the tangent or $\chi_{\mathbf{h}}$ formulas as far as navigating random phases toward their solution values and avoiding false minuma. But the degree of R_2 convergence can be improved by late-stage $\chi_{\mathbf{h}}$ refinement and difference mapping to resolve overlapped atomic sites, which incidentally do not add significantly to the total computation time normally required. This presentation is not meant to imply that the standard R_1 target function should be totally abandoned in preference to R_2, as R_1 may still have some advantage to move the initial atom-based values of the phases away from local false minima that the R_2 function might otherwise gravitate toward. We do not suggest that R_2-based $\chi_{\mathbf{h}}$ refinement and difference mapping should be performed in every refinement cycle, as they are undoubtedly more valuable as an end-stage SnB refinement strategy to drive the final set of phases toward a stable convergent minimum. At this point, for any solution, since the majority of the correct atomic sites are located near the top of the peak list, peak-list optimization (Sheldrick & Gould, 1995) could be effectively used to help eliminate more of the spurious entries at the end of the last SnB refinement cycle. The analysis performed here was not abetted by efforts either to artificially extrapolate the limited data sets to higher resolution (Jia-xing *et al.*, 2005; Caliandro *et al.*, 2005*a*,*b*), or extend the basis set of triples to other types of invariants or improve their reliability estimates prior to phasing, as these would tend to mask the effectiveness of the modifications that were performed. Those methods remain viable with regard to further extending the improvements described in this report.

We thank Melda Tugac and Gloria Del Bel for their assistance in this project. Drs V. Z. Pletnev, S. Parkin, S. E. Ealick, A. P. Kuzin, L. Tong and D. Ghosh kindly provided the high-resolution data sets for the macromolecular test cases described. Research support from the Human Frontier Science Program grant (HFSP-RGP0021/2006-C) is gratefully acknowledged. We appreciate the collaborative efforts and encouragement of Professors Alberto Podjarny, Nobuo Niimura and Peter Timmins in this regard.

References

Bunkóczi, G., Vértesy, L. & Sheldrick, G. M. (2005). *Angew. Chem. Int. Ed.* **44**, 1340–1342.
Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Spagna, R. (2005). *J. Appl. Cryst.* **38**, 381–388.
Burla, M. C., Giacovazzo, C. & Polidori, G. (2011). *J. Appl. Cryst.* **44**, 193–199.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*a*). *Acta Cryst.* D**61**, 556–565.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*b*). *Acta Cryst.* D**61**, 1080–1087.
Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2009). *J. Appl. Cryst.* **42**, 302–307.
Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. A. (1997). *Acta Cryst.* A**53**, 436–444.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.
Ghosh, D., Erman, M., Sawicki, M., Lala, P., Weeks, D. R., Li, N., Pangborn, W., Thiel, D. J., Jörnvall, H., Gutierrez, R. & Eyzaguirre, J. (1999). *Acta Cryst.* D**55**, 779–784.
Hauptman, H. A. (1988). *Proceedings of the American Crystallographic Association Meeting*, Philadelphia, USA, Abstract R4, p. 53.
Hauptman, H. A., Xu, H., Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* A**55**, 891–900.
Jia-xing, Y., Woolfson, M. M., Wilson, K. S. & Dodson, E. J. (2005). *Acta Cryst.* D**61**, 1465–1475.
Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.

Miller, R., Galitsky, N. M., Duax, W. L., Langs, D. A, Pletnev, V. Z. & Ivanov, V. T. (1993). *Int. J. Peptide Protein Res.* **42**, 539–549.
Pal, A., Debreczeni, J. É., Sevvana, M., Gruene, T., Kahle, B., Zeeck, A. & Sheldrick, G. M. (2008). *Acta Cryst.* D**64**, 985–992.
Pletnev, V. Z., Galitskii, N. M., Smith, G. D., Weeks, C. M. & Duax, W. L. (1980). *Biopolymers*, **19**, 1517–1534.
Pletnev, V. Z., Ivanov, V. T., Langs, D. A., Strong, P. D. & Duax, W. L. (1992). *Biopolymers*, **32**, 819–827.
Pletnev, V. Z., Mihailova, I. Y., Ivanov, V. T, Langs, D. A, Grochulski, P. & Duax, W. L. (1991). *Biopolymers*, **31**, 409–415.
Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nat. Methods*, **6**, 651–653.
Rollett, J. S. (1965). *Computing Methods in Crystallography*, pp. 35–37. Oxford: Permagon Press.
Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.
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.
Tong, L., Warren, T. C., King, J., Betageri, R., Rose, J. & Jakes, S. (1996). *J. Mol. Biol.* **256**, 601–610.
Usón, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Xu, H. & Hauptman, H. A. (2004). *Acta Cryst.* A**60**, 153–157.
Xu, H. & Hauptman, H. A. (2006). *Acta Cryst.* D**62**, 897–900.
Xu, H., Hauptman, H. A. & Weeks, C. M. (2002). *Acta Cryst.* D**58**, 90–96.