Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 30 November 1999
Accepted 17 January 2000

Shake-and-Bake applications using simulated reference-beam data for crambin

Charles M. Weeks,[a]* Hongliang Xu,[a] Herbert A. Hauptman[a] and Qun Shen[b]

[a]Hauptman–Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203, USA, and [b]Cornell High Energy Synchrotron Source (CHESS), 283 Wilson Laboratory, Cornell University, Ithaca, NY 14853, USA. Correspondence e-mail: weeks@hwi.buffalo.edu

The *Shake-and-Bake* method, as implemented in the computer program *SnB*, has been applied to simulated reference-beam data for the small protein crambin at several resolutions in the range 1.5–3.0 Å. Sets of triplet invariants were generated having simulated mean triplet-phase errors from 0 to 60°. Provided that these errors were no larger than 40°, it was possible (at all resolutions tested) to find trial sets of individual Bragg phases with mean errors of 40–45°. At 1.5 Å, this could be achieved using only a single reference-beam data set. Peak picking provided useful phase constraints even at the lowest resolution tested. These results suggest that direct methods may be useful in conjunction with reference-beam data at resolutions lower than 1.2Å.

1. Introduction

Shake-and-Bake (Weeks *et al.*, 1994) is a multisolution or multitrial direct-methods procedure that alternates reciprocal-space phase refinement with peak picking in real space to impose constraints through a physically meaningful interpretation of the electron density. Phase refinement can utilize either the tangent formula (Karle & Hauptman, 1956) or the technique of parameter shift (Bhuiya & Stanley, 1963) to reduce the value of the minimal function (Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.* 1994). Although the *Shake-and-Bake* approach has increased, by an order of magnitude, the size of structures solvable by direct methods (Deacon *et al.*, 1998), these successes have been limited, with few exceptions, to structures for which the diffraction data can be measured to at least 1.2 Å. The hirustasin structure, which can be determined using 1.55 Å truncated data, currently holds the record for the lowest-resolution successful application of direct methods to a complete structure (Usón *et al.*, 1999). On the other hand, 3 Å isomorphous or anomalous difference data are sufficient to permit successful *Shake-and-Bake* applications to large substructures such as the 70 site selenomethionine derivative of a 370 kDa epimerase enzyme (Deacon *et al.*, 1999).

The so-called triplet structure invariants,

$$\Phi_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \tag{1}$$

where the φs are the phases of the corresponding structure factors, provide the foundation for phase-determining relationships, such as the tangent formula or the minimal function, used in direct methods. Successful applications rely on the use of accurate probabilistic estimates, provided by the Cochran (1955) distribution, for the values of the triplet invariants and their corresponding cosines (Germain *et al.*, 1970). Simulation experiments have shown that the structure of the small protein crambin can be solved by *Shake-and-Bake* even at 2 Å if the invariants used are accurate enough (Weeks *et al.*, 1998). Therefore, the primary breakdown of *Shake-and-Bake* during low-resolution applications seems to occur in reciprocal space, and such failures could probably be overcome if a sufficient number of accurate invariant values were available.

Recent work in the field of multiple-beam diffraction provides grounds for hope that a general method for experimentally measuring phase information can be found. For example, it has been shown that triplet phases,

$$\delta_{\mathbf{HG}} = -\varphi_{\mathbf{H}} + \varphi_{\mathbf{G}} + \varphi_{\mathbf{H}-\mathbf{G}}, \tag{2}$$

can be measured for lysozyme with a mean error of approximately 20° (Weckert *et al.*, 1993; Weckert & Hümmer, 1997). In addition, direct methods strengthened by simulated known triplet phases have been used to redetermine the structure of rubredoxin at 1.54 Å (Mo *et al.*, 1996) as well as the structure of bovine pancreatic trypsin inhibitor at resolutions as low as 2 Å (Mathiesen & Mo, 1997, 1998). Unfortunately, the one-at-a-time methods currently used to measure triplet phases seriously limit practical applications. However, the recently proposed reference-beam diffraction method (Shen, 1998, 1999), in which a single Bragg reflection (**G**) serves as a reference beam and is common to many simultaneously recorded triplet phases, would permit large numbers of these phases to be measured quickly.

As illustrated in Fig. 1, the geometry of the reference-beam experiment is a simple conceptual modification of the direct-beam geometry used in the conventional oscillation camera set-up. Instead of being perpendicular to the incident X-ray beam, the oscillation axis is tilted by the Bragg angle ($\theta_{\mathbf{G}}$) of a strong reference reflection ($\mathbf{G}$) that is oriented to coincide with the oscillation axis. In this way, reflection $\mathbf{G}$ can be fully excited throughout the crystal oscillation or rotation. The intensities of all Bragg reflections recorded on an area detector during such an oscillation will be affected by the existence of the $\mathbf{G}$-reflected wave ($\mathbf{k_G}$), which is coherently split from the incident wave ($\mathbf{k_0}$) and can be viewed as a new incident wave. Thus, $\mathbf{k_G}$ can produce its own diffracted beams during an oscillation. Therefore, for each Bragg reflection ($\mathbf{H}$) excited by the original incident beam ($\mathbf{k_0}$), there exists another reflection ($\mathbf{H} - \mathbf{G}$) excited by $\mathbf{k_G}$, whose wavevector ($\mathbf{k_{H-G}}$) is parallel to the original $\mathbf{k_H}$ of the $\mathbf{H}$ reflection. The two sets of diffraction patterns, one excited by $\mathbf{k_0}$ and the other by the reference beam $\mathbf{k_G}$, coincide in space and interfere with each other, producing a phase-sensitive image on the area detector.

The three-beam interference between the diffracted wave for the $\mathbf{H}$ reflection and the wave diffracted through reflections $\mathbf{G}$ and $\mathbf{H} - \mathbf{G}$ depends on the relative phase difference [$\delta_{\mathbf{HG}}$ of equation (2)], which is the triplet phase measured in the reference-beam experiment. By applying Friedel's law and changing variables, it is easy to show that the triplet structure invariants ($\Phi_{\mathbf{HK}}$) of direct methods and the triplet phases ($\delta_{\mathbf{HG}}$) of multiple-beam diffraction are equivalent. Individual estimates or measurements of triplet invariant values ($\omega_{\mathbf{HK}}$) can be accommodated by a modified tangent formula,

$$\tan\varphi_{\mathbf{H}} = \frac{\sum_{\mathbf{K}} W_{\mathbf{HK}} \sin(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} W_{\mathbf{HK}} \cos(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})}, \qquad (3)$$

or by a modified minimal function,

$$R(\Phi) = \left(2\sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}}\right)^{-1} \sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}}\{[\cos(\Phi_{\mathbf{HK}}) - \cos(\omega_{\mathbf{HK}})]^2 + [\sin(\Phi_{\mathbf{HK}}) - \sin(\omega_{\mathbf{HK}})]^2\} \qquad (4)$$

(Weeks *et al.*, 1998). The $W_{\mathbf{HK}}$ are appropriately chosen weights and the $\Phi_{\mathbf{HK}}$ are computed from the current values of the trial phases. Either of these relationships can serve as the basis for a modified *Shake-and-Bake* procedure.

Since all invariants measurable in a single reference-beam experiment have a common reflection ($\mathbf{G}$), it is important to learn whether such an invariant set contains sufficient information that it could provide the basis for a successful *Shake-and-Bake* application. This question was addressed in a series of experiments that were conducted using simulated reference-beam data and are reported here. The effects of limited resolution, as well as errors in the triplet phases, were also examined. Finally, the results of dual-space (*Shake-and-Bake*) refinement based on these invariant sets were compared to the results of reciprocal-space phase refinement alone (conventional direct methods).

2. Materials and methods

Reference-beam measurements were simulated using the 0.83 Å intensity data and refined phases for crambin, a 46 residue protein containing 327 unique non-H atoms as well as the equivalent of about 75 fully occupied water molecules (Teeter *et al.*, 1993). The data were truncated to several different resolutions (1.5, 2.0, 2.5 and 3.0 Å), and the remaining reflections with the largest structure-factor magnitudes ($|F|$) were chosen for use as simulated reference-beam reflections ($\mathbf{G}$s). Among the seven reflections (724, 020, 404, $64\bar{1}$, $20\bar{5}$, 044 and $40\bar{5}$) used as reference beams, the one with highest resolution (3.56 Å) was 044. All triplet invariants involving each $\mathbf{G}$ were then generated provided that $F_{\mathbf{H}}$ and $F_{\mathbf{H-G}}$ were greater than three times their corresponding standard deviations. Next, 'error-free' values of the corresponding triplet phases were computed using the known values of the individual phases. Subsequently, a random-number generator was used to assign uniformly distributed errors to the triplet phases in such a way that sets of triplets were created having mean triplet-phase errors of 10, 20, ..., 60°. Finally, sets of invariants involving more than one $\mathbf{G}$ reflection were constructed by concatenating the desired number of reference-beam data sets. Table 1 specifies the number of invariants used and the number of reflections that could be phased with different numbers of reference-beam data sets at each of the resolutions studied.

Phasing experiments were carried out with version 2.0 of the computer program *SnB* (Weeks & Miller, 1999*a*) altered to use the modified tangent formula [equation (3)] as the means for phase refinement. In these experiments, the simulated

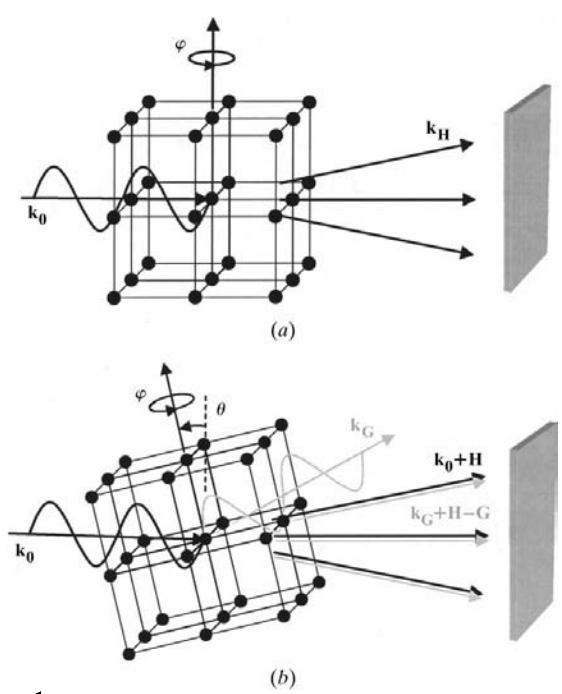

Figure 1
Comparative representations of (*a*) direct-beam and (*b*) reference-beam geometry. In the reference-beam set-up, two sets of diffraction patterns (black and gray) interfere and create a phase-sensitive image.

Table 1
Numbers of invariants used and reflections phased at each resolution with different numbers of reference-beam data sets.

The total numbers of unique reflections that exist are 5543, 2391, 1238 and 734 at resolutions of 1.5, 2.0, 2.5 and 3.0 Å, respectively.

	1.5 Å resolution		2.0 Å resolution		2.5 Å resolution		3.0 Å resolution	
Data sets	Refl.	Inv.	Refl.	Inv.	Refl.	Inv.	Refl.	Inv.
1	5421	6725	2316	2591	1188	1125	681	544
2	5239	10626	2341	4217	1221	1902	719	956
3	5247	14537	2348	5834	1227	2664	725	1358
4	5250	21324	2351	8458	1231	3820	731	1917
5	5250	25095	2352	9994	1231	4531	731	2285
6	–	–	2352	12525	1231	5615	731	2800
7	–	–	–	–	–	–	731	3145

triplet-phase values were used as $\omega_{\mathbf{HK}}$ and all weights ($W_{\mathbf{HK}}$) were taken to be unity. Using standard *Shake-and-Bake* protocol, 1000 initial trial structures were created, each of which consisted of 125 randomly positioned atoms. Each trial structure was refined at each of the four resolutions using error-free triplet phases as well as the sets of triplet phases having different amounts of random error. In each case, two different phasing experiments were conducted. In the first set of experiments, 100 cycles of conventional direct-methods phase refinement were performed in reciprocal space alone. The other set of experiments involved 100 cycles of dual-space (*Shake-and-Bake*) refinement in which modified-tangent phase refinement was alternated with selection of the 125 largest peaks. The choice of 125 peaks for these experiments was based on the observation that this is close to the optimum number of selected peaks even for 0.83 Å data (Weeks & Miller, 1999*b*) and it was expected that this number might decrease at lower resolutions because fewer atoms were expected to be clearly distinguished.

Solutions were unequivocally identified on the basis of mean phase error (*i.e.* lowest mean phase difference from the correct phases for some choice of origin and enantiomorph). Trials with mean phase errors less than 50° were counted as 'solutions' and the 'success rate' for each experiment was defined as the percentage of trial structures that refined to solutions. The quantity

$$m(\Phi) = 1 - \left(\sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}}\right)^{-1} \sum_{\mathbf{H},\mathbf{K}} W_{\mathbf{HK}} \cos(\Phi_{\mathbf{HK}} - \omega_{\mathbf{HK}}), \qquad (5)$$

where $\Phi_{\mathbf{HK}}$ are the triplet phases computed using the *SnB* refined phases and the $\omega_{\mathbf{HK}}$ are the measured triplet phases, can be computed without prior knowledge of the true phases, and it was examined as a potential figure of merit.

3. Results

The results of the simulation experiments described above are summarized in Fig. 2. First, it is apparent that, for crambin, *Shake-and-Bake* solutions are obtainable from simulated reference-beam data, even at a resolution as low as 3 Å. At 1.5 Å, one reference-beam data set (curve G1 in Fig. 2*e*) is sufficient, but more data sets are required as the resolution decreases. The success rate decreases as the mean triplet-phase error increases, and the maximum tolerable mean triplet-phase error is approximately 50°. The maximum tolerable mean error also tends to increase as the number of data sets increases but to decrease as the resolution decreases. In several cases, the success rate was fairly constant until the mean error approached 40° but then it decreased rapidly.

It is also clear that the success rate is higher, and the number of required data sets less, if the dual-space refinement technique (*Shake-and-Bake*) is used rather than refinement in reciprocal space alone. This is true even at the lowest resolution tested (3 Å). Thus, it appears that using peak picking to place density in approximately correct positions has a beneficial effect as a phase constraint, even at resolutions this low. Not unexpectedly, the average distance between each peak and the nearest true atomic position increased as resolution decreased. Although the effects of varying the number of peaks selected were not studied in detail, several experiments were conducted with the G3 invariant set at 2 Å. Success rates were found to be similar when 125, 150 or 200 peaks were selected for inclusion in subsequent structure-factor calculations, but they were significantly reduced when only 100 peaks were selected. Further study is needed to determine the optimum number of peaks as a function of resolution.

The lowest values of the final mean individual phase error were in the range of 40–45°, regardless of whether reciprocal-space or dual-space refinement was used. Since relatively low values (0.60–0.68) of $m(\Phi)$ [equation (5)] were found to be strongly correlated with low values of the mean phase error at all resolutions examined, $m(\Phi)$ appears to be a potentially useful figure of merit. When error-free triplet phases were used, the lowest values of the mean phase error were in the range 40–45°; when the mean triplet-phase random errors were in the range 30–40°, the best mean phase errors were only slightly larger (46–49°). In general, $m(\Phi)$ demonstrated greater discriminatory power when it was applied to a larger number of reference-beam data sets or higher-resolution data. There was a clear distinction (*i.e.* a bimodal distribution) between the mean phase errors for solutions (<50°) and nonsolutions (>70°) at each of 1.5 and 2 Å. On the other hand, the mean phase-error distribution was continuous (*i.e.* unimodal) at 2.5 Å unless the G6 invariant set was used, and it was continuous for all G values at 3 Å. The $m(\Phi)$ distributions were also unimodal in most of the cases tested. However, $m(\Phi)$ still served as an effective figure of merit in the sense that, at any resolution, trial phase sets with relatively small average errors could be identified if the number of reference beams was chosen to be sufficiently large.

4. Conclusions

These *Shake-and-Bake* experiments involving simulated reference-beam data for crambin have demonstrated that final phase values having average errors in the range 40–45° can be obtained *ab initio* using the *SnB* program. Furthermore, invariants having simulated average triplet-phase errors even as large as 50° can be tolerated. Although it appears that more than one reference reflection will be required when the resolution is 2 Å or less, the required number of such reflections is not large. In higher-resolution cases, there is hope that a single reference beam will provide sufficient information despite the fact that all triplet invariants in the set share a common Bragg reflection. Furthermore, the actual mean triplet phase error (20°) observed for multiple-beam measurements for lysozyme is less than the maximum simulated error (40–50°) used successfully in the current study. In addition, an effective figure of merit has been found that identifies phase sets with low mean phase error and, therefore, distinguishes solutions from nonsolutions. It is also especially noteworthy that the peak-picking procedure used in the real-space segment of the *SnB* program is effective at resolutions as low as 3 Å provided that the triplet phase values used in the reciprocal-space segment are sufficiently accurate.

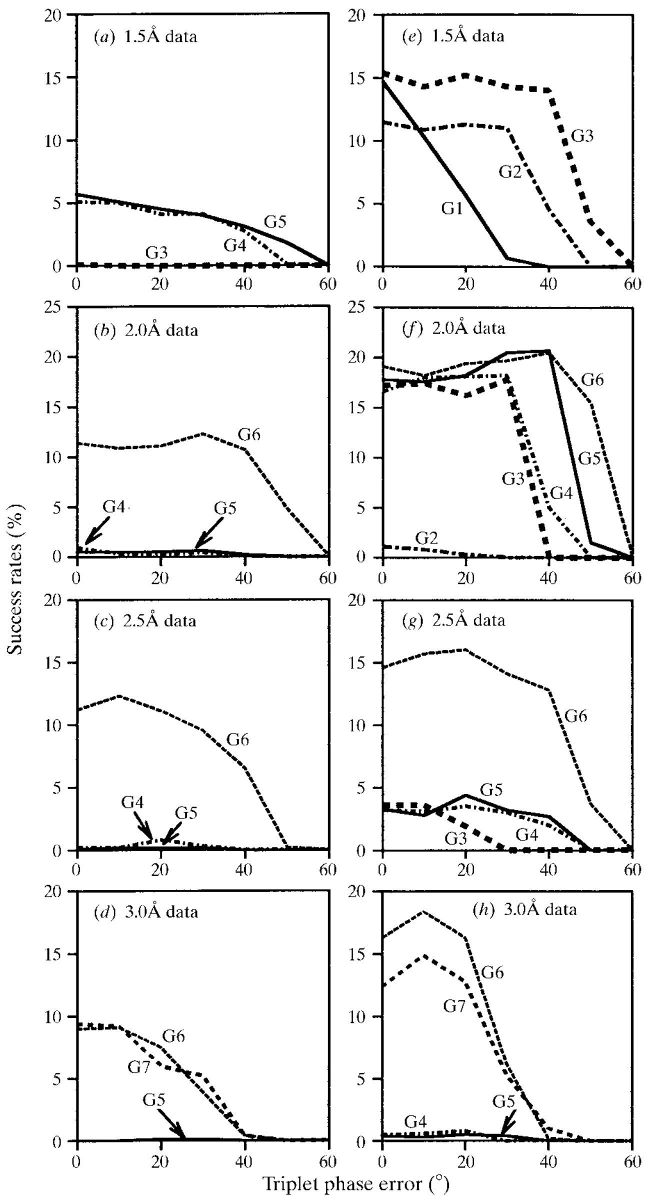

Figure 2
SnB success rates using simulated reference-beam data for crambin. (*a*)–(*d*) show the results of modified tangent refinement in reciprocal space alone (conventional direct methods). (*e*)–(*h*) illustrate the results of dual-space (*Shake-and-Bake*) refinement. The labels G1 through G7 indicate the number of reference-beam data sets used. Comparison of (*a*)–(*d*) with the corresponding (*e*)–(*h*) figures clearly shows the superior performance of *Shake-and-Bake* relative to the conventional direct methods.

We would like to express our appreciation to Melda Tugac for preparing the figures. This research was supported by NIH grant GM-46733. CHESS is supported by NSF grant DMR-9713424.

References

Bhuiya, A. K. & Stanley, E. (1963). *Acta Cryst.* **16**, 981–984.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
Deacon, A. M., Ni, Y., Coleman, W. G. Jr & Ealick, S. E. (1999). Am. Crystallogr. Assoc. Annual Meeting, Buffalo, NY, USA, Abstract PT21.
Deacon, A. M., Weeks, C. M., Miller, R. & Ealick, S. E. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.
Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Germain, G., Main, P. & Woolfson, M. M. (1970). *Acta Cryst.* B**26**, 274–285.
Hauptman, H. A. (1991). *Crystallographic Computing 5: from Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 324–332. IUCr/Oxford University Press.
Karle, J. & Hauptman, H. A. (1956). *Acta Cryst.* **9**, 635–651.
Mathiesen, R. H. & Mo, F. (1997). *Acta Cryst.* D**53**, 262–268.
Mathiesen, R. H. & Mo, F. (1998). *Acta Cryst.* D**54**, 237–242.
Mo, F., Mathiesen, R. H., Hauback, B. C. & Adman, E. T. (1996). *Acta Cryst.* D**52**, 893–900.
Shen, Q. (1998). *Phys. Rev. Lett.* **80**, 3268–3271.
Shen, Q. (1999). *Phys. Rev. B*, **59**, 11109–11112.
Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). *J. Mol. Biol.* **230**, 292–311.
Usón, I., Sheldrick, G. M., de la Fortelle, E., Bricogne, G., di Marco, S., Priestle, J. P., Grütter, M. G. & Mittl, P. R. E. (1999). *Structure*, **7**, 55–63.
Weckert, E. & Hümmer, K. (1997). *Acta Cryst.* A**53**, 108–143.
Weckert, E., Schwegle, W. & Hümmer, K. (1993). *Proc. R. Soc. London Ser. A*, **442**, 33–46.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Weeks, C. M. & Miller, R. (1999*a*). *J. Appl. Cryst.* **32**, 120–124.
Weeks, C. M. & Miller, R. (1999*b*). *Acta Cryst.* D**55**, 492–500.
Weeks, C. M., Miller, R. & Hauptman, H. A. (1998). In *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 463–468. Dordrecht: Kluwer Academic Publishers.