Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 1 May 2000
Accepted 29 June 2000

A hybrid minimal principle for the crystallographic phase problem

Xiangan Liu and Wu-Pei Su*

Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, TX 77204, USA. Correspondence e-mail: wpsu@uh.edu

Simulated annealing is used to solve the X-ray phase problem formulated as a minimization problem. The cost function consists of two parts, one represents the discrepancy between measured and calculated intensities while the other monitors the probability distribution of the triplets. From a random real-space structure at the start, the atoms are moved one by one to gradually reduce the cost function until the best structure emerges. Trial calculations for structures including hexadecaisoleucinomycin (HEXIL) are presented. Comparison of this method with other related methods is made.

1. Introduction

We have been pursuing a real-space approach (Su, 1995; Giacovazzo, 1998) to the X-ray phase problem. An initial random structure is gradually modified until it settles into one for which the calculated intensities best fit the observed intensities. Simulated annealing (Kirkpatrick *et al.*, 1983) is used for this procedure to avoid being trapped in local minima of the cost function. Although the method has enjoyed some success (Chen *et al.*, 1997; Wang *et al.*, 1999), it requires extensive diffraction data and the computer time grows dramatically with the size of the molecule. The origin of this difficulty could probably be traced to some peculiar nature of the landscape of the cost function and the phase-transtion-like behavior in the annealing curve. To speed up the algorithm, we have attempted to modify the cost function by adding another piece to it. This additional cost is the one used in the 'Shake-and-Bake' (SnB) method (Miller *et al.*, 1993). It measures the deviation of the triplet structural invariants from the theoretical distribution. It turns out that such an addition has significantly improved the convergence of the original real-space algorithm.

From a different perspective, this hybrid minimal principle shares some similarity with SnB as both real space and reciprocal space are involved. This probably explains why it appears to be an improvement over the original pure real-space approach. There are, however, marked differences between the two methods. Whereas the atomicity and positiveness of the electron-density function are automatic in our method, they have to be enforced by the laborious procedures of Fourier synthesis and peak picking in the SnB method.

In the following, we first explain the new methodology and then describe several examples, followed by a discussion. As a new method, it might offer distinct advantages for solving certain structures.

2. Methodology

In the original real-space approach, the phase problem is formulated as a minimization problem of the following cost function:

$$R_1(\{\mathbf{r}_i\}, \lambda) = \sum_{\mathbf{k}} [\lambda |F(\mathbf{k})| - |F(\mathbf{k})|_{\rm obs}]^2, \qquad (1)$$

where the structure factor $F(\mathbf{k})$ is calculated from the atomic scattering factor $f_i(\mathbf{k})$ and atomic coordinates $\mathbf{r}_i$ in the usual fashion.

$$F(\mathbf{k}) = \sum_i f_i(\mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{r}_i), \qquad (2)$$

the summation being over all N atoms in the unit cell. Like the particle coordinates $\mathbf{r}_i$, the scale factor λ is determined by the minimization procedure.

The above formalism obviously remains valid upon replacing the $F(\mathbf{k})$ by the normalized structure factors $E_{\mathbf{k}}$. In addition, an arbitrary weight can be given to each term in the summation in (1).

$$R_1(\{\mathbf{r}_i\}) = \sum_{\mathbf{k}} w(E_{\mathbf{k}})(|E_{\mathbf{k}}| - |E_{\mathbf{k}}|_{\rm obs})^2. \qquad (3)$$

Our past experience with the above algorithm indicates that a successful structure determination within a reasonable amount of computer time usually requires a very high ratio of the number of reflections over the number of independent atoms. This ratio reaches about 240 in a $P1$ structure containing 75 atoms. A smaller ratio seems to imply a flat landscape for the cost function near the true minimum and therefore demands many more Monte Carlo steps to locate it accurately.

We have attempted to alleviate the above problem by adding the following minimum-variance structure-invariant residual (DeTitta *et al.*, 1994) to R_1:

$$R_2(\varphi) = \alpha \sum_{\mathbf{h},\mathbf{k}} w(A_{\mathbf{hk}})[\cos(\varphi_{\mathbf{h}} - \varphi_{\mathbf{k}} + \varphi_{\mathbf{k}-\mathbf{h}}) - I_1(A_{\mathbf{hk}})/I_0(A_{\mathbf{hk}})]^2, \quad (4)$$

where $A_{\mathbf{hk}} = 2|E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{k}-\mathbf{h}}|N^{-1/2}$ and N is the number of non-H atoms in the primitive reduced cell (Langs *et al.*, 1995). α is a scale to adjust the overall weight of R_2 relative to R_1. For the particular choice of weight $w(A_{\mathbf{hk}}) = A_{\mathbf{hk}}$, R_2 reduces to the minimal function used in the well known SnB method.

In the evaluation of (4), *observed* magnitudes of the normalized structure factors and phases of the *calculated* structure factors are used. Therefore, in each update of the atomic coordinates, there is on average an improvement in both the magnitude and the phase of the calculated structure factors. This makes the Monte Carlo step more discriminative and accelerates the convergence toward the true minimum.

The choice of the weight w in (3) and (4) is quite flexible. We have chosen them to favor larger $|E_{\mathbf{k}}|$ and $A_{\mathbf{hk}}$, respectively. α is chosen so that R_1 is comparable to R_2. After the modification of the cost function, the annealing procedure follows pretty much that employed previously in a pure real-space approach (Su, 1995; Giacovazzo, 1998). We only indicate some minor changes in the algorithm. At each annealing temperature, each atom is updated a certain number of times. Instead of moving the atom in a random direction, we move the atoms in one of the x, y and z directions each time. The step size of the movement starts typically at a few tenths of the shortest side of the unit cell and decreases gradually with annealing temperature. In addition, we impose the constraint that interatomic distance should be larger than 1.1 Å.

In the following, we report several trial calculations. As a representative symmetry group, we focus exclusively on the $P2_12_12_1$ space group with $Z = 4$.

3. Examples

3.1. Virginiamycin ($C_{43}H_{49}N_7O_{10}\cdot 3CH_3OH$)

The structure was originally solved by a version of *MULTAN* (Declercq *et al.*, 1978). We treat all the non-H atoms as C atoms. From the synthetic reflection data of 1 Å, we pick the largest 1000 E's to form the real-space cost R_1. For R_2, we compose 3584 major triplets from 842 of the 1000 reflections. At each annealing temperature, every atom is updated 2000 times in each of the three directions.

To understand the formation of the correct structure at low annealing temperature, we examine the annealing curve. As in both pure real-space and pure phase approaches, this formation is signalled by the appearance of a phase-transition-like feature (Chen *et al.*, 1997; Chen & Su, 2000) in the annealing curve. We are refering to the sharp drop of the cost function at about $T = 78$ in Fig. 1. It is also instructive to plot each component of the cost function R_1 and R_2 separately, as shown in Fig. 2. It seems that below the transition temperature R_1 continues to drop and eventually reach zero, as it should for a completely correct solution. R_2 on the other hand saturates at a non-zero limit.

At the lowest annealing temperature, 166 out of the 167 special phases are correct. Altogether, 700 out of the 842 phases are correct within 15°. Viewed in real space, 64 out of the 66 nonhydrogen atoms (Fig. 3) are in their correct positions.

The entire calculation took a few CPU hours on a Digital 500 MHz Alpha Workstation. We have successfully reproduced the structure in several independent calculations, so the algorithm seems to be quite efficient and robust.

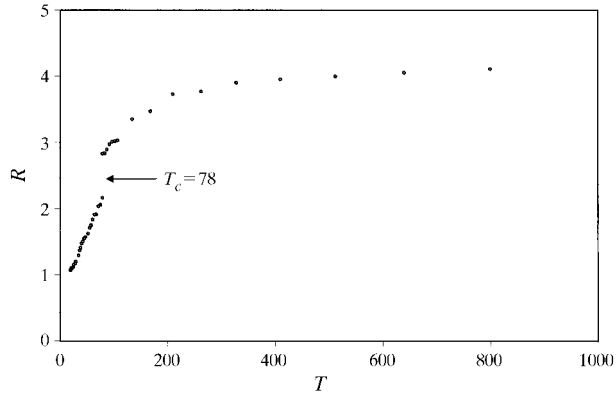

Figure 1
An annealing curve (total residual *versus* temperature) that yields the correct structure for virginiamycin.

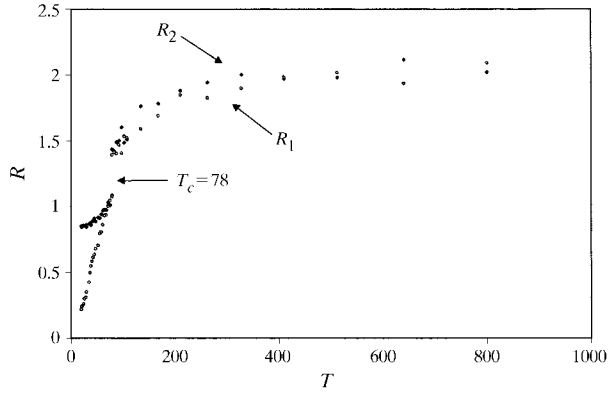

Figure 2
Same as in Fig. 1, with the costs R_1 and R_2 plotted separately.

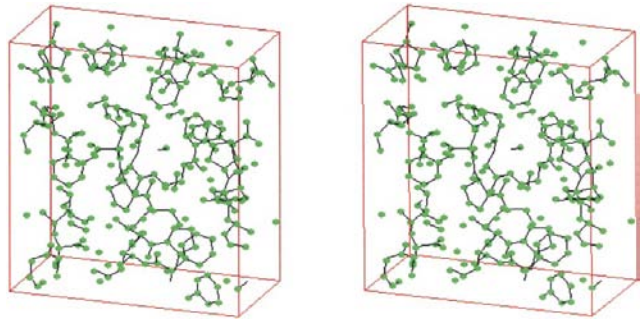

Figure 3
Stereoview of the calculated structure compared with the native structure of virginiamycin in wire frame.

3.2. Isoleucinomycin ($C_{60}H_{102}N_6O_{18}$)

We have picked this structure (Pletenev *et al.*, 1980) and HEXIL to facilitate comparison with the SnB method. From the actual diffraction data, 1950 E's are used to construct R_1 and among them 766 are involved in making up 4018 triplets for R_2. We initially treat all the atoms as C atoms to determine the center of mass of the atoms.

One run that leads to the correct structure exhibits the annealing behavior depicted in Fig. 4. There is a sudden drop of the total residual at temperatures around 74. It is instructive to look at the evolution of the phases as the system approaches the transition region. At point A on the annealing curve, 95 out of the 176 special phases are correct, which is no better than a random set of phases. At point B, however, 149 special phases are correct and 319 out of the total 766 phases are correct within 0.1π. At point C, essentially all special phases (175 out of 176) are right. 490 of the 766 phases are accurate within 0.1π. In real space, this implies that 76 out of the 84 atoms are in their correct positions (within 0.1 Å). A completely correct structure can be obtained by further straightforward real-space refinement with more reflections. The nitrogen and oxygen atoms can be recovered by further peak-height optimization (Su, 1995).

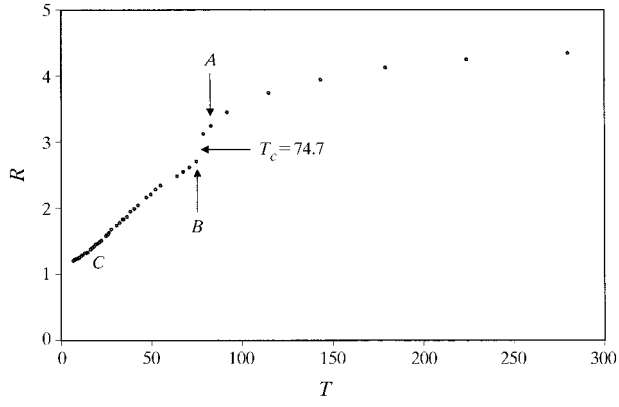

Figure 4
An annealing curve for isoleucinomycin.

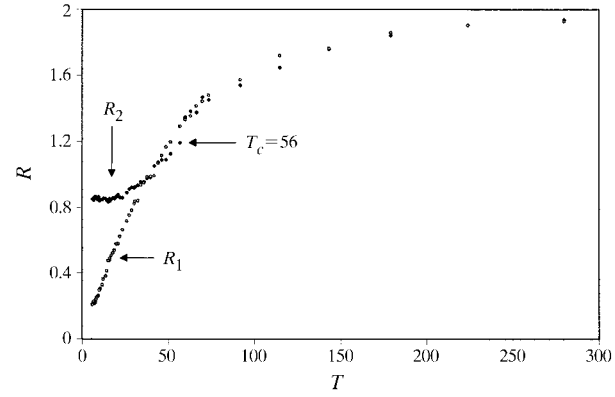

Figure 5
An annealing curve for HEXIL. R_1 and R_2 are plotted separately.

The entire run took about 8 h CPU on the Alpha Workstation. As in the previous example, the result is easily reproduced in separate independent runs.

3.3. HEXIL ($C_{80}H_{136}N_8O_{24}$)

This structure was solved by Pletenev *et al.* (1992). We drop the water molecules and treat all non-hydrogen atoms as C atoms in fabricating the diffraction data. The 2000 largest reflections and 4174 triplet structural invariants are included in the cost function.

In a particular run, the cost functions R_1 and R_2 are plotted as a function of the annealing temperature in Fig. 5. The two curves track each other closely at temperatures above $T = 40$. Below that, R_2 saturates at about 0.9 whereas R_1 continues to head towards zero. R_1 can converge to zero because in the trial calculations the data are synthetic error-free data; in contrast, R_2 saturates because the target values $I_1(A_{\mathbf{hk}})/I_0(A_{\mathbf{hk}})$ are only probabilistic estimates, not exact. Although there is still a recognizable phase-transition-like feature in the annealing curve, it is much less pronounced than in the previous examples. The final structure obtained is displayed in Fig. 6 as balls superimposed on the native structure in the wire frame. More than 90% of the atoms fall on the correct positions. As in the case of isoleucinomycin, a completely correct solution can be obtained by refinement in real space with more reflection data.

Some more specifics of the calculation are as follows: At each temperature, each atom is updated 3000 times, the maximum step size of the atomic displacement decreases with temperature; it is about 4 Å at $T = 70$. The entire calculation described above took about one week of CPU time on the 500 MHz Alpha Workstation. The correct structure was actually pretty much established after two days at about

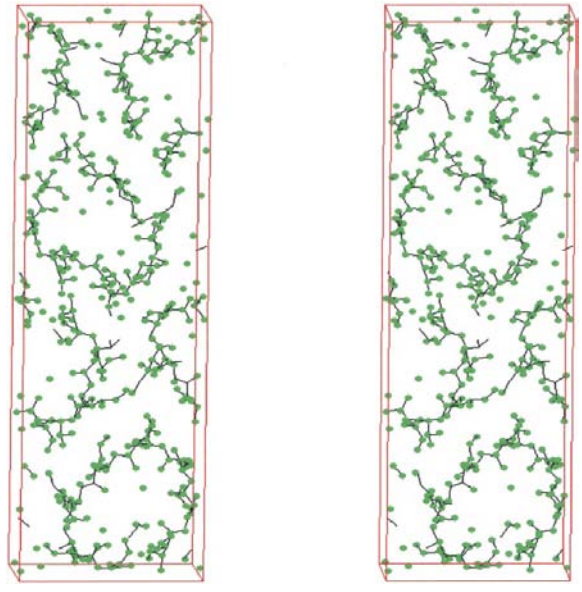

Figure 6
Stereoview of the calculated structure of HEXIL superimposed on the native structure in wire frame.

$T = 50$. We have repeated the calculation; the success rate is nearly 100%.

We have not been able to solve the same structure with the same amount of data in a pure real-space approach. The cost function R_1 drops significantly with temperature as shown in Fig. 7, yet no sensible structure is visible even at the lowest temperature. A comparison with the R_1 curve in Fig. 5 reveals a nearly constant difference in cost function at low temperature. This suggests that the system is trapped in a glassy state in a pure real-space approach.

4. Discussion

Many of the successful direct methods to date involve computation in both real and reciprocal spaces. In the hybrid minimal principle presented above, we have achieved real-space filtering without peak peaking. This has considerably simplified the algorithm. As a result, it requires a fairly modest amount of computer time to solve the structures presented.

The relative weight of R_2 with respect to R_1 signifies the importance of improving the phase as compared to the improvement in the magnitude of the structure factors. It is a parameter that can be fine-tuned to maximize the efficiency of the algorithm. Empirically, we find that the optimal ratio is about 1:1.

For isoleucinomycin, we have resolved the structure using both real and synthetic diffraction data for comparison. The ideal data make some but not much difference in the efficiency of the algorithm. For HEXIL, it does seem to make a more genuine difference; after a few attempts, we still have not succeeded in solving the structure using the real data.

As in our previous minimization calculations, annealing is essentially here to avoid being trapped in a local mimimum. Although the phase-transition-like feature seems to become less conspicuous as the structure becomes larger, it is still there. Annealing hardly yields the correct structure unless one starts with a temperature above the transition temperature.

We believe that larger structures can be solved given more computer time. A closer comparison of the present method with the Shake-and-Bake method is of obvious interest. Such benchmark calculations are reserved for a future publication.

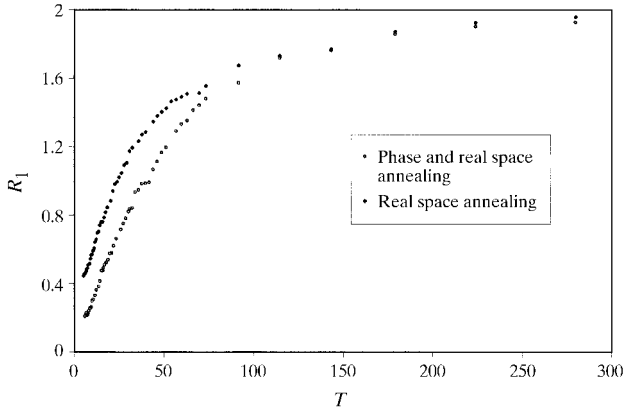

Figure 7
An annealing curve for HEXIL in a pure real-space approach. The R_1 curve in Fig. 5 is reproduced here for easy comparison.

This work was partially supported by the Texas Center for Superconductivity, the Texas Advanced Research Program under grant no. 003652-707-1997, the Texas Advanced Technology Program under grant no. 003652-0222-1999 and the Robert A. Welch Foundation. We gratefully acknowledge D. A. Langs for providing the diffraction data of isoleucinomycin and HEXIL and for useful correspondence.

References

Chen, Y. & Su, W.-P. (2000). *Acta Cryst.* A**56**, 127–131.
Chen, Y.-S., Su, W.-P., Mallela, S. P. & Geanangel, R. A. (1997). *Acta Cryst.* A**53**, 396–399.
Declercq, J. P., Germain, G., Van Meerssche, M., Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* B**34**, 3644–3648.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Giacovazzo, C. (1998). *Direct Phasing in Crystallography*, pp. 198–201. Oxford University Press.
Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. (1983). *Science*, **220**, 671–680.
Langs, D. A., Miller, R., Hauptman, H. A. & Han, G. Y. (1995). *Acta Cryst.* A**51**, 81–87.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Pletenev, V. Z., Galitskii, N. M., Smith, G. D., Weeks, C. M. & Duax, W. L. (1980). *Biopolymers*, **19**, 1517–1534.
Pletenev, V. Z., Ivanov, V. T., Langs, D. A., Strong, P. & Duax, W. L. (1992). *Biopolymers*, **32**, 819–827.
Su, W.-P. (1995). *Acta Cryst.* A**51**, 845–849.
Wang, X., Chen, Y.-S. & Su, W.-P. (1999). *J. Appl. Cryst.* **32**, 409–412.