Improved Monte Carlo sampling in a real space approach to the crystallographic phase problem

Xiangan Liu and W. P. Su
Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, Texas 77204
(Received 20 August 2002; published 16 December 2002)

A real space approach has been proposed to solve the x-ray phase problem formulated as a minimization problem. The cost function consists of two parts: one represents the usual crystallographical residual while the other enforces the probability distribution of the invariant phase triplets. Starting from a random real space structure, the atoms are moved one by one to gradually reduce the cost function (simulated annealing). In addition, the atoms are encouragd to preferentially sample the high density regions in space determined by an approximate density map which in turn is updated and modified by averaging and Fourier synthesis. Such a reduction of the configurational space has led to considerable improvement of the algorithm compared to an earlier version. Trial calculations for structures including hexadecaisoleucinomycin (HEXIL) and a collagen-like peptide (PPG) are presented.

DOI: 10.1103/PhysRevE.66.066703 PACS number(s): 02.70.Tt, 61.10.−i, 42.30.Rx

I. INTRODUCTION

We have been pursuing a real space approach to the x-ray phase problem. The idea is to generate a real space configuration of the atoms so that the calculated intensities best match the observed diffraction data [1], or equivalently the discrepancy (the crystallographical residual) between calculated and observed intensities is minimized. A Monte Carlo updating can be used to gradually improve an initial random configuration so that hopefully a correct structure emerges at the end of this simulated annealing procedure [2]. We have learned over the years that the usual crystallographical residual is a very effective indicator only when the molecular configuration is already close to the correct structure. In other words, the radius of convergence of this minimization method is quite small, especially for large molecules. As a remedy, we found [3] that supplementing the residual with another indicator (the Hauptman minimal function) can substantially increase the radius of convergence. Basically, the Hauptman function [4] monitors the improvement in the phases of the structure factors as the atoms move in real space. In this paper, we show that the algorithm can be further improved by maintaining an approximate density map during the entire annealing procedure. The atoms are encouraged to sample preferentially the high density regions according to the approximate density map. This map is constantly updated and modified through averaging and Fourier synthesis.

To facilitate the comparison with our previous work, we report on two familiar structures, the isoleucinomycin and the HEXIL, both in the $P2_12_12_1$ geometry. Solution of HEXIL with real diffraction data has been a challenge for us in the past. With the new methodology, it can be easily achieved on a PC. The same happens to a triple-helix peptide (PPG). We anticipate similar speedup with larger structures.

An important ingredient of the new methodology is the reduction of the configurational space sampled by a trial atomic displacement. In a typical Metropolis updating scheme [5], an atom is moved to a random new test position. In contrast, the move is made more selective or smarter by reference to an existing approximate density map in the current algorithm. The selective sampling idea has been proposed by one of us [6] to refine a low resolution structure into a progressively higher resolution one. To implement the reduction, Chou and Lee [7] have recently considered a guided simulated annealing method which we incorporate in this paper.

II. METHODOLOGY

We have refered to a particular version of the real space approach as a hybrid minimum principle [3] because the total cost function consists of two pieces, the crystallographical residual (R_1) and the minimum-variance structural-invariant residual (R_2, the Hauptman function). The Hauptman function measures the deviation of a triplet of phases from the theoretical expectation value. R_1 and R_2 are given by

$$R_1(\{\vec{r}_i\}) = \sum_{\mathbf{k}} w(E_{\mathbf{k}})(|E_{\mathbf{k}}| - |E_{\mathbf{k}}|_{obs})^2, \tag{1}$$

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

where $A_{\mathbf{hk}} = 2|E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{k}-\mathbf{h}}|N^{-1/2}$, I_1 and I_0 are the modified Bessel functions, and N is the number of non-H atoms in the primitive reduced cell [8]. α is a scale to adjust the overall weight of R_2 relative to R_1. In the evaluation of Eq. (2), *observed* magnitudes of the normalized structure factors $E_{\mathbf{k}}$ and phases $\varphi_{\mathbf{k}}$ of the *calculated* structure factors are used. The structure factors are Fourier transforms of the electron density, which is a function of the atomic coordinates $\{\vec{r}_i\}$. The choice of the weight w in Eqs. (1) and (2) is quite flexible. We have chosen them to favor larger $|E_k|$ and A_{hk}, respectively. α is chosen so that R_1 is comparable to R_2.

Minimization of the total cost with respect to the atomic coordinates yields the correct structure in principle when enough diffraction data is available. Simulated annealing is used in the minimization procedure. We refer to previous expositions of simulated annealing [9,10] for details. A central step in the annealing is the choice of a new trial atomic position. A common practice is to adopt the Metropolis

scheme which picks a new position randomly. In a biased Monte Carlo step, which we employ here, that choice is done with reference to an approximate density map. The map is obtained from a current molecular configuration by centering a sphere of radius 1.1 Å at each atom. In selecting a new test position for each atom, a high probability is given if the test position lies within one of the spheres. As the annealing proceeds and progressively better molecular configurations are obtained, the reference density map is updated accordingly.

The method outlined above solves the isoleucinomycin structure easily. For HEXIL and PPG, however, additional procedure of averaging and Fourier synthesis are needed for the solution. They are described in the next section.

III. EXAMPLES

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

The choice of parameter values and real diffraction data [11] are described in our previous work [3]. The only new addition here is the way to choose a new trial position. As described above, this is implemented simply by giving a higher probability (90%) to a position which lies inside as opposed to outside of an approximate density map. The density map has been constructed by centering a sphere around each atom in a molecular configuration which has the lowest total cost so far in the annealing process. The approximate density map is updated after we have moved all the atoms once, provided that the new molecular configuration is better than the old one in terms of the total cost.

An annealing cycle which begins with a random structure at high temperature and ends with a better structure at low temperature takes about 15 CPU minutes on a single Pentium IV processor. This is a short run compared to our previous calculation because at a given annealing temperature, each atom is updated only fifty times. Not every cycle yields the correct structure. The success rate is about one in three.

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

This has been a difficult structure [12] for us [3,13] previously. With real diffraction data, we have not been able to solve it using the hybrid minimum principle. Neither have we succeded in solving it using the annealing method described above.

To proceed further, we superimpose the low temperature molecular configurations from ten independent cylces. The choice of origin and enantiomorph is determined as usual by the phases of four specially chosen reflections. To obtain a clear profile of the molecules, we keep only atoms whose ''density'' exceed a certain value. Here the ''density'' of an atom is defined by the number of atoms within a distance of 1.5 Å from it. In other words, we keep only atoms which are surrounded by sufficiently large number of atoms nearby (within 1.5 Å). Figure 1 is an example of what the remaining atoms look like in the unit cell. They are depicted by the dots. The wire frame is the native structure for comparison.

From such a composite molecular configuration, we place a small sphere (of radius 0.5 Å) at the center of each atom. Using this as a reference density map, we go through another annealing cycle and end up with a single low temperature molecular structure. During this and subsequent annealing cycles, for simplicity we do not update the reference density map. With the phases calculated from this structure and the observed diffraction intensity, we carry out a Fourier synthesis to construct a smooth electron density function. The high density region defined by this density function is then used as a reference density map for the next round of annealing and Fourier cycle. After ten such cycles, we end up with a structure shown in Fig. 2, where about half of the atoms are in correct positions. This structure can be refined into a completely correct structure by adding more reflections to the cost function and by going through a regular annealing (without biased Monte Carlo).

The entire calculation can be accomplished on a single Pentium IV processor in about ten CPU hours, which is a modest amount of CPU time considering the difficulty of the structure.

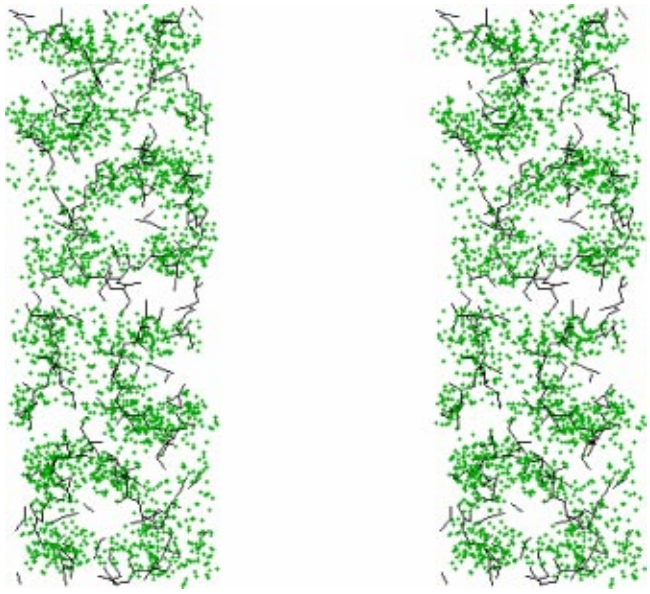

FIG. 1. (Color online) Stereoview of a reference density map (the dots) obtained by averaging over the results of ten preliminary runs. The wire frame is the native HEXIL structure.

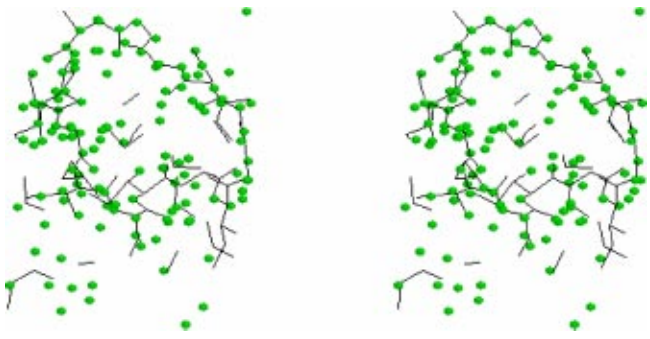

FIG. 2. (Color online) Stereoview of the calculated structure of HEXIL. The wire frame is the native structure of HEXIL. For clarity, only a quarter of the unit cell is shown.

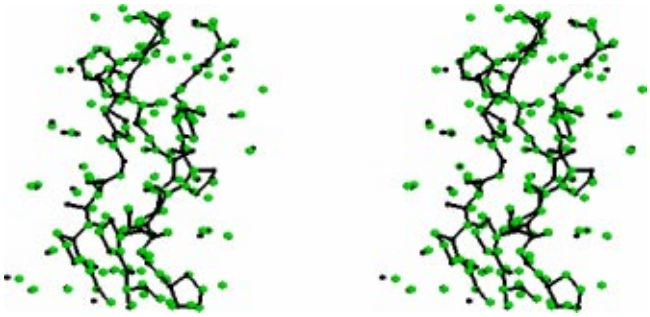

FIG. 3. (Color online) Stereoview of the calculated structure of the PPG triple-helix peptide superimposed on the native structure in wire frames. Only a quarter of the unit cell is shown.

C. Triple-helix peptide (PPG)

This is a collagenlike peptide [14] with the repeating sequence (Pro-Pro-Gly). The Protein Data Bank code is 1g9w. Diffraction data are fabricated from the atomic coordinates provided by the PDB file. All the 126 peptide atoms and the 36 water molecules are treated as carbons. The number of reflections and the number of triplets included in the cost function are comparable to those of HEXIL. The space group is also $P2_12_12_1$.

This structure being considerably larger than the two previous ones, it requires more extensive averaging. Within each annealing cycle, we retain the best configuration for each annealing temperature. Below a certain temperature, all the configurations are superimposed to yield the first average. The second average is done over independent cycles. The final density is then subjected to Fourier synthesis and other modifications to serve as the approximate density map for further annealing, as described before.

After about five iterations of extensive averaging and Fourier synthesis, there is a dramatic enhancement in the quality of the density map. About 70% of the atoms (the gray dots; green dots online only) are in their correct positions in the final calculated molecular configuration shown in Fig. 3. Such a configuration can be easily refined into a completely correct structure.

The entire calculation takes about one day of CPU time on a single Pentium IV processor.

IV. DISCUSSION

We have been approaching the crystallographical phase problem from the real space side [1,2]. The simplest form of which is to extract the atomic coordinates directly from the intensity data by a least squares fit, i.e., by minimizing R_1. Annealing is used to avoid being trapped in a false minimum. To enlarge the radius of convergence, we have deviated from a pure real space approach and have included a fit to the theoretical probability distribution of the structure-invariant phases [3]. Biased sampling is further employed here to reduce the configurational space making the sampling more effective. By doing so we end up with a simple and yet powerful algorithm which can be straightforwardly applied to a typical problem. It is particularly appropriate for structures having simple motifs such as the ring in the HEXIL or the helix in the PPG peptide. Since large structures such as proteins usually contain motifs in the form of alpha helices and beta sheets, they might be amenable to our treatment.

As Chou and Lee have adopted a similar approach in their work [7], it is instructive to make a comparison. First, our construction of the approximate density function is quite different from theirs. We do not go through a multiresolution procedure, but we modify the density with Fourier synthesis. Second, we incorporate the phase information in seeking a better atomic position. Both features contribute to make our algorithm more efficient.

Finally, it is worthwhile pointing out that despite the relevance of the phase information in our algorithm, the laborious process of peak picking which is required in the Shake-and-Bake method [4], is not needed because we keep track of atomic positions throughout the calculations.

ACKNOWLEDGMENTS

This work was partially supported by the Texas Center for Superconductivity, the Texas Advanced Technology Research 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 the HEXIL and for useful correspondence. Thanks are also due to C. I. Chou and T. K. Lee for sending us their report.

[1] J. Karle, Proc. Natl. Acad. Sci. U.S.A. **88**, 10099 (1991).
[2] W.P. Su, Acta Crystallogr., Sect. A: Found. Crystallogr. **A51**, 845 (1995).
[3] X. Liu and W.P. Su, Acta Crystallogr., Sect. A: Found. Crystallogr. **A56**, 525 (2000).
[4] R. Miller, G.T. DeTitta, R. Jones, D.A. Langs, C.M. Weeks, and H.A. Hauptman, Science **259**, 1430 (1990).
[5] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[6] W.P. Su, Physica A **221**, 193 (1995).
[7] C.I. Chou and T.K. Lee, Acta Crystallogr., Sect. A: Found. Crystallogr. **A58**, 42 (2002).
[8] D.A. Langs, R. Miller, H.A. Hauptman, and G.Y. Han, Acta Crystallogr., Sect. A: Found. Crystallogr. **A51**, 81 (1995).
[9] S. Kirkpatrick, C.D. Jr. Gelatt, and M.P. Vecchi, Science **220**, 671 (1983).
[10] C. Giacovazzo, *Direct Phasing in Crystallography* (Oxford University Press, New York, 1998).
[11] V.Z. Pletnev, N.M. Galitskii, G.D. Smith, C.M. Weeks, and W.L. Duax, Biopolymers **19**, 1517 (1980).
[12] V.Z. Pletnev, V.T. Ivanov, D.A. Langs, P. Strong, and W.L. Duax, Biopolymers **32**, 819 (1992).
[13] Y. Chen and W.P. Su, Acta Crystallogr., Sect. A: Found. Crystallogr. **A57**, 733 (2001).
[14] L. Vitagliano, R. Berisio, L. Mazzarella, and A. Zagari, Biopolymers **58**, 459 (2001).