Structure determination of organic molecules from diffraction data by simulated annealing

Hsiao-Ping Hsu,[1] Ulrich H. E. Hansmann,[2] and Simon C. Lin[1,3,*]

[1]*Computing Centre, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[2]*Department of Physics, Michigan Technological University, Houghton, Michigan 49931-1291*
[3]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*

We study simulated annealing techniques for crystal structure determination from diffraction data. We demonstrate that for this problem the efficiency of simulated annealing can be systematically improved by an iterative simulation protocol. Our approach is tested for the example of 9-(methylamino)-1 H-phenalen-1-one-1, 4-dioxan-2-yl hydroperoxide solvate ($C_{18}H_{19}NO_5$).

I. INTRODUCTION

Crystal structure resolution from powder diffraction data is often the only way to characterize the structural properties of macromolecules. This is because many compounds cannot be grown as single crystals and exist only in the form of microcrystalline powders. While the information extracted from single-crystal and powder diffraction patterns is essentially the same, the technical difficulties in extracting the three-dimensional structure from the one-dimensional diffraction patterns are much more severe in the case of powder diffraction data. In the past, most research focused on inorganic compounds, and here many structures could be solved from powder diffraction data by direct methods [1–5] and Patterson method [6–12]. However, due to weaker scattering intensity and poor diffraction data at high angles, these methods often fail for organic compounds. As a consequence, only few structures of organic compounds could be solved from powder diffraction data.

The above described technical difficulties led to an ongoing search for more sophisticated numerical techniques and for extracting structural properties of organic compounds from powder diffraction data. One successful attempt in that direction is the development of algorithms that work in the direct space (as opposed to the phase space). Common to these techniques is that they attempt to minimize the differences between observed and calculated diffraction patterns by moving and rotating single atoms or even the whole molecular fragments. In that way, the problem of structure prediction is translated into a global optimization problem. The Monte Carlo approach [13], genetic algorithms [14], simulated annealing [15–17], and grid search [18–20] are some examples of the optimization techniques used for structure prediction from diffraction data.

Among the above examples, simulated annealing (SA) [21,22] is probably the most established stochastic optimization technique. In the past few years, SA has been used successfully for the traveling salesman problem [23], designing complex integrated circuits, image processing, etc. [24]. Its application to crystal structure resolution of small molecules is well documented [15–17]. In this paper, we evaluate the

usefulness of SA for structure determination of organic macromolecules from diffraction data. We show that the efficiency of simulated annealing can be systematically improved by an iterative simulation protocol. Our approach is tested for the case of 9-(methylamino)-1 H-phenalen-1-one-1, 4-dioxan-2-y1 hydroperoxide solvate ($C_{18}H_{19}NO_5$) [25]. The paper is organized as follows. In Sec. II we first briefly review the details of our simulation approach, the results are presented and discussed in Sec. III. Finally, we present our conclusions in Sec. IV.

II. METHODS

Use of simulated annealing for structure prediction from diffraction data requires definition of a proper cost function. In our case, we try to rearrange positions of atoms in order to minimize the difference between calculated and observed diffraction patterns. Hence, our cost function can be defined by

$$E = \frac{\sum_{j=1}^{N_k} (|F_{cal}(\mathbf{k}_j)| - |F_{obs}(\mathbf{k}_j)|)^2}{\sum_{j=1}^{N_k} |F_{obs}(\mathbf{k}_j)|^2} \times 100, \tag{1}$$

where N_k is the total number of reflections, $F_{obs}(\mathbf{k}_j)$ is the observed structure factor, and $F_{cal}(\mathbf{k}_j)$ is the calculated structure factor. The structure factor is expressed as

$$F(\mathbf{k}_j) = \sum_{i=1}^{N} f_i \exp(i 2\pi(\mathbf{k}_j \cdot \mathbf{x}_i)), \tag{2}$$

where N is the total number of atoms in a unit cell, $\mathbf{x}_i$ are the fractional coordinates of the ith atom. The scattering factor of an atom (the so called atomic form factor) f_i is given by

$$f_i(\sin\theta/\lambda) = \sum_{q=1}^{4} a_{qi} \exp(-b_{qi}\sin^2\theta/\lambda^2) + c_i, \tag{3}$$

where a_{qi}, b_{qi}, and c_i are the coefficients of the ith atom and can be obtained from Ref. [26]. Note that due to periodic boundary conditions and the symmetry of the space group only the atoms in an asymmetric unit need to be considered.

*Corresponding author. Email address: sclin@sinica.edu.tw

The convergence of the diffraction pattern, generated by simulated annealing, towards the observed pattern is usually also monitored by a so-called residual value, or R factor. This quantity is defined by

$$R = \frac{\sum_{j=1}^{N_k} \left\| F_{\text{obs}}(\mathbf{k}_j) \right| - \left| F_{\text{cal}}(\mathbf{k}_j) \right\|}{\sum_{j=1}^{N_k} \left| F_{\text{obs}}(\mathbf{k}_j) \right|}, \qquad (4)$$

and serves a measure of the precision of the refinement results.

Simulated annealing generates a sequence of temperatures $(T_1, T_2, T_3, \ldots, T_{N_T})$ with a cooling rate α, i.e.,

$$T_i = \alpha T_{i-1} \quad \text{and} \quad T_1 > T_2 > T_3 > \cdots > T_{N_T}. \qquad (5)$$

The initial temperature T_1 is chosen such that the acceptance ratio is about 0.5, and the last temperature T_{N_T} has to be chosen sufficiently low. Note that the total number of temperatures N_T in the annealing process has to be chosen large in order that the cooling rate

$$\log \alpha = (\log T_{N_T} - \log T_1)/(N_T - 1) \qquad (6)$$

is slow enough for the system to stay in (quasi) equilibrium at any stage of the annealing approach. For every tempera-
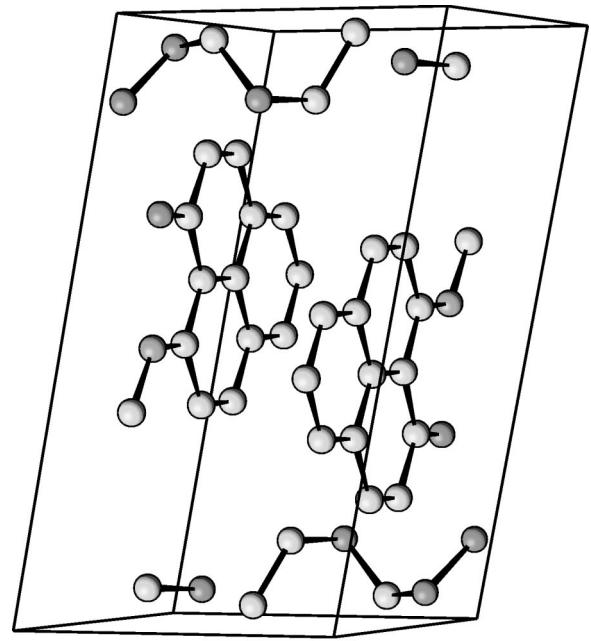

FIG. 1. The crystal structure of $C_{18}H_{19}NO_5$ in a unit cell.

ture T_i, N_S Monte Carlo sweeps are performed. A Monte Carlo sweep is a sequence of M Metropolis steps, with M the number of atoms in an asymmetric unit. In every Metropolis step one tries to change the position of an atom according to

TABLE I. The minimal values of average energy $\bar{E}$ and average R factor $\bar{R}$ for the three cases $N_S = 2048$, $N_S = 20\,480$, and $N_S = 102\,400$ without switch and the two cases $N_S = 2048$ and $N_S = 20\,480$ with switch.

| | Without switch | | | | | | With switch | | | |
| | $N_S = 2048$ | | $N_S = 20\,480$ | | $N_S = 102\,400$ | | $N_S = 2048$ | | $N_S = 20\,480$ | |
No.	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$
1	25.75	0.53	16.66	0.42	4.83	0.22	22.74	0.50	0.39	0.07
2	6.10	0.25					19.76	0.46	0.37	0.07
3	20.45	0.46	4.94	0.23			16.51	0.43	0.39	0.07
4	24.29	0.51	4.33	0.21			20.05	0.47	19.65	0.46
5	18.16	0.44	4.34	0.21			17.16	0.42	0.38	0.07
6	18.42	0.45	4.28	0.21			16.42	0.42	0.38	0.07
7	17.83	0.44	16.76	0.42	4.88	0.22	17.48	0.44	16.26	0.41
8	7.99	0.26					17.16	0.43	16.42	0.42
9	16.69	0.42	5.08	0.23			0.39	0.07		
10	17.11	0.42	17.42	0.43	5.63	0.24	16.48	0.41	0.39	0.07
11	18.92	0.44	16.74	0.42	4.36	0.21	0.41	0.07		
12	5.73	0.24					0.38	0.07		
13	19.53	0.46	19.13	0.45	3.70	0.20	16.50	0.43	16.47	0.41
14	17.52	0.43	4.91	0.22			0.36	0.07		
15	6.07	0.25					18.92	0.45	16.44	0.42
16	19.07	0.45	3.75	0.20			19.40	0.45	16.42	0.42
17	16.16	0.41	17.56	0.43	3.53	0.19	20.64	0.46	0.38	0.07
18	21.48	0.48	16.62	0.42	4.53	0.22	16.92	0.43	0.38	0.07
19	5.12	0.23					18.68	0.44	0.38	0.07
20	20.15	0.46	6.45	0.25			16.94	0.43	0.40	0.07
%	25		65		100		20		70	
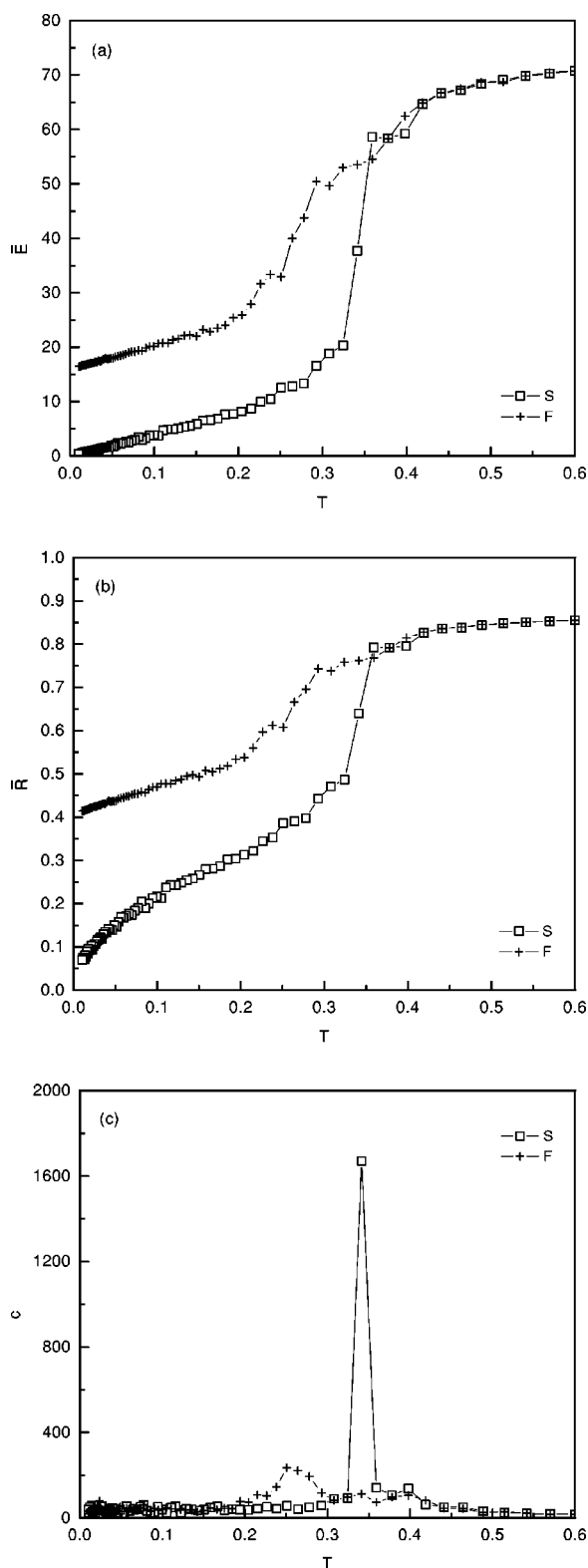
$$x'^l = x^l + \Delta x^l. \qquad (7)$$

Here, x^l, $l = 1$, 2, and 3, are the fractional coordinates of the given atom, $\Delta x^l = r_s^l \eta_l$ is the displacement in the direction l, r_s^l is a scale factor to ensure equal acceptance ratios in three directions, and η_l is a random number between $+0.5$ and -0.5. Such a proposed move of an atom is then accepted with a probability $\min[1, \exp(-\Delta E/T)]$, and otherwise rejected.

III. RESULTS AND DISCUSSIONS

In this paper we focus on structure prediction of organic molecules from diffraction data. Our test system is 9-(methylamino)-1 H-phenalen-1-one-1, 4-dioxan-2-y1 hydroperoxide solvate ($C_{18}H_{19}NO_5$). This molecule has the space group $P\bar{1}$ (Triclinic), and the cell constants are $a = 6.9520$ Å, $b = 9.6900$ Å, $c = 12.5410$ Å, $\alpha = 77.11°$, $\beta = 73.78°$, and $\gamma = 80.62°$. The number of formula units per cell is $Z = 2$. Its structure was solved by direct methods (DIRDIF [27]) and refined (on F^2) using SHELXL-93 [28] with anisotropic thermal parameters for all non-hydrogen atoms. Hydrogens were located from difference maps and refined isotropically. The crystal structure of this molecule is shown in Fig. 1. Since we are in this paper mainly concerned with methodological questions, we decided to use synthetic data instead of original experimental pattern. For this purpose we reconstructed the diffraction pattern for $C_{18}H_{19}NO_5$ from the positions of the known structure [25] using POWDERCELL software [29]. In this way we obtain a 2θ(angle)-I(intensity) diagram. The software also provides methods to give an index (h,k,l) to each reflection that is extracted from the ''perfect'' experimental result. The hydrogen atoms are ignored and all calculations rely on a set of 2823 reflections.

Our simulations are started from random configurations in which the positions of the 24 non-hydrogen atoms in an asymmetric unit are chosen randomly. The remaining atoms in a unit cell are generated by applying the symmetry operations of the space group. Utilizing Eq. (2), the structure factor is calculated by using all the atoms in a unit cell. Note that the structure factor can be simplified to

$$F(\mathbf{k}_j) = 2 \sum_{i=1}^{N/2} f_i \cos[2\pi(\mathbf{k}_j \cdot \mathbf{x}_i)], \qquad (8)$$

where only half the atoms needed to be calculated, since the molecule is centrosymmetric in this case.

In all our SA runs, the initial temperature is chosen to be 0.6 and a sequence of $N_T = 80$ temperatures is generated with cooling rate $\alpha = 0.95$. For every temperature we perform N_S Monte Carlo sweeps, and we compared the three cases $N_S = 2048$, $N_S = 20\,480$, and $N_S = 102\,400$. The computational time need for such SA run on a Pentium II PC with 450 MHz was 4.42 hr ($N_S = 2048$), 43.50 hr ($N_S = 20\,480$), and 188.69 hr ($N_S = 102\,400$), respectively. Our analyses relies on 20 such SA run starting from 20 different random start configu-

FIG. 2. (a) Average energy $\bar{E}$, (b) average R factor $\bar{R}$, and (c) the specific heat c as functions of T by simulated annealing for a successful run (S) and one where SA failed to find the correct structure (F). The results rely on $N_S = 2048$ Monte Carlo sweeps for every temperature.

TABLE II. The minimal values of average energy $\bar{E}$ and average R factor $\bar{R}$ and the number of iterations n for the case $N_S^1 = 2048$.

No.	Without switch								With switch					
	$n=1$		$n=2$		$n=3$		$n=4$		$n=1$		$n=2$		$n=3$	
	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$	$\bar{E}$	$\bar{R}$
1	25.75	0.53	4.89	0.22					22.74	0.50	0.38	0.07		
2	6.10	0.25							19.76	0.46	19.43	0.46	0.40	0.07
3	20.45	0.46	3.67	0.20					16.51	0.43	0.37	0.07		
4	24.29	0.51	2.98	0.18					20.05	0.47	0.37	0.07		
5	18.16	0.44	18.29	0.44	20.13	0.46	5.63	0.24	17.16	0.42	0.37	0.07		
6	18.42	0.45	16.80	0.42	6.75	0.26			16.42	0.42	0.37	0.07		
7	17.83	0.44	5.47	0.23					17.48	0.44	0.44	0.07		
8	7.99	0.26							17.16	0.43	0.36	0.07		
9	16.69	0.42	3.00	0.18					0.39	0.07				
10	17.11	0.42	4.30	0.21					16.48	0.41	0.38	0.07		
11	18.92	0.44	6.20	0.25					0.41	0.07				
12	5.73	0.24							0.38	0.07				
13	19.53	0.46	5.03	0.23					16.50	0.43	0.38	0.07		
14	17.52	0.43	4.25	0.21					0.36	0.07				
15	6.07	0.25							18.92	0.45	0.37	0.07		
16	19.07	0.45	4.34	0.21					19.40	0.45	17.27	0.43	0.41	0.07
17	16.16	0.41	4.95	0.22					20.64	0.46	0.36	0.07		
18	21.48	0.48	5.66	0.24					16.92	0.43	16.73	0.43	0.38	0.07
19	5.12	0.23							18.68	0.44	16.22	0.42	0.40	0.07
20	20.15	0.46	20.15	0.46	6.06	0.25			16.94	0.43	20.69	0.48	0.38	0.07
%	25		85		95		100		20		75		100	

rations. This allows us to calculate from Eqs. (1) and (4) the average energy (cost function) $\bar{E}$, the average R factor $\bar{R}$, and the specific heat c defined by

$$c = \beta^2 (\overline{E^2} - \bar{E}^2), \qquad (9)$$

with the inverse temperature defined by $\beta = 1/kT$ and $k=1$.

We now describe our results in the following. By the formulation of our cost function, the correct structure has energy $E=0$ and a residual value $R=0$. However, we found that a less stringent criteria can be used to determine whether the correct structure was found. We have checked that all configurations with a value of $R<0.3$ resemble closely to the (known) crystal structure of our molecule. For this reason, we choose $R<0.3$ as criterion to identify the correct crystal structure. Table I lists our results for the three cases. In the first case ($N_S = 2048$) the probability to find the correct structure is only 25%, but increasing the number of sweeps at each temperature by a factor 10 to $N_S = 20\,480$ raises that probability to 60%. Further enlarging the statistics to $N_S = 12\,400$ allows us to find the correct crystal structure in 100% of all the runs.

Our results can be further improved by introducing an additional global update that takes into account that the values of atomic form factors of atoms C, N, and O are close to each other and difficult to distinguish in a SA run. For this reason, we choose for every temperature N_p pairs of atoms and switch the positions of the atoms in every pair once. If $\Delta E < 0$, the switch is accepted, otherwise the switch is rejected. The results are also listed in Table I. While this additional move affects little the probability of finding the target structure, it decreases the R value in the cases where the structure was found, i.e., leads to more refined structures. We remark that the results were worse when switches were accepted or rejected through the Metropolis algorithm.

While the above results show that in principle simulated annealing is able to find the correct structure, they also demonstrate the limitations of such a simple cooling protocol. The necessary number of Monte Carlo sweeps is not known, *a priori*. In order to improve the efficiency of our method we further investigated the behavior of $\bar{E}$ and c as functions of T. Figure 2 displays these quantities for two typical cases, one where SA failed to find the correct structure and one where SA converged. In both cases, the number of sweeps was set to $N_S = 2048$. Comparing the two cases, we see that the case where the correct structure was found there is a sudden drop in the average energy $\bar{E}$ as the temperature T decreases (and a corresponding maximum in the specific heat c). This step-function-like behavior of the energy is missing in all cases where SA failed to find the correct structure.

The above observation motivates an iterative SA method controlled by the specific that we propose now. The method is described as follows:

(1) We start with an initial SA run, in which we set the number of Monte Carlo sweeps to the same value N_S for every temperature. Through the SA run we measure the spe-
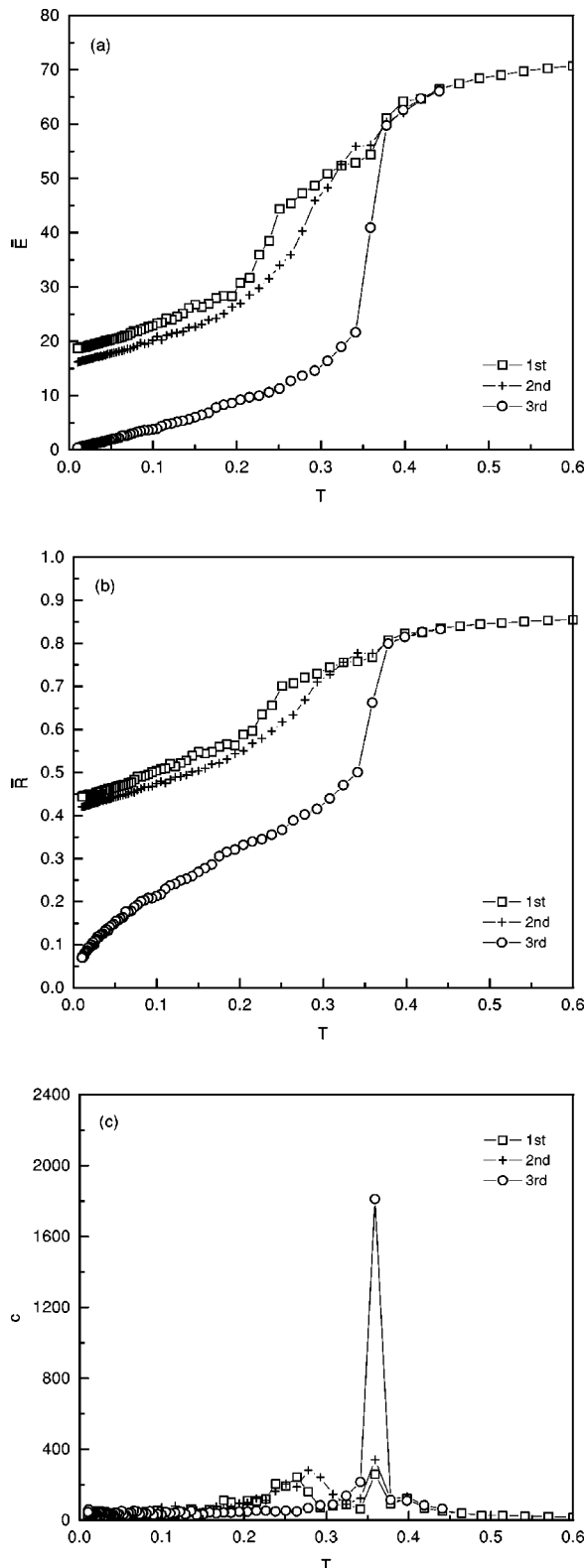
FIG. 3. (a) Average energy $\bar{E}$, (b) average R factor $\bar{R}$, and (c) specific heat c as functions of T and the number of iterations in our iterative simulated annealing procedure.

cific heat c as a function of T. If the SA run converges (i.e., $R<0.3$), we stop and accept the last configuration as our crystal structure.
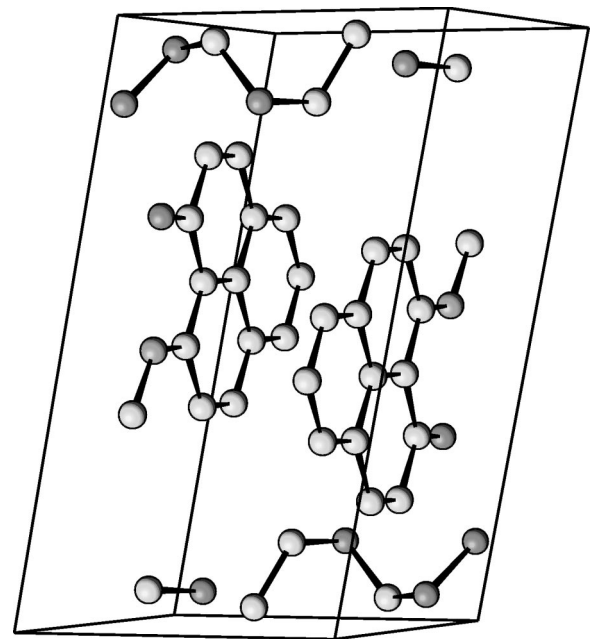
(2) If the SA does not converge, we now set at each tem-

FIG. 4. The crystal structure of $C_{18}H_{19}NO_5$ in a unit cell as determined by our iterative annealing protocol.

perature the number of sweeps $N_S(T)$ to

$$N_S^{n+1}(T)=N_S^n(T)\cdot c_n(T)/c_n(T_1),\qquad(10)$$

with T_1 a high temperature.

(3) A new SA run is now started with these new values of $N_S(T)$. If it finds the correct structure, then the process stops otherwise the iterative process is repeated until the target configuration (for example, defined by $R<0.3$ criteria) is reached.

We have employed this approach to all trapped samples in our previous simulations, and the methods performed very well. The results are shown in Table II. A sample that presents iterative improvement is shown in Fig. 3. The figure documents that there is a sudden drop in the average energy at the third SA run and the shape of the specific heat changes successively from a multiple-maxima form to a single sharp maximum. Figure 4 displays the corresponding lowest-energy structure that has a R value of $R=0.07$.

IV. CONCLUSIONS

We have applied simulated annealing techniques to crystal structure determination of organic compounds from diffraction data. Our results show the feasibility of such an approach. Using a simple iterative annealing protocol, which is controlled by values of the specific heat, we find the correct structure of our test molecule in all cases. This observation leads to the hope that such a refinement of the annealing protocol may allow the use of SA for crystal structure prediction even from powder diffraction data or in the case of much larger molecules (such as proteins).

ACKNOWLEDGMENTS

[1] H.A. Hauptman, Science **233**, 178 (1986).

[2] M.M. Woolfson, Acta Crystallogr., Sect. A: Found. Crystallogr. **A43**, 593 (1987).

[3] *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. Ladd and R. A. Palmer (Plenum Press, New York, 1980).

[4] R. Miller, G.T. DeTitta, R. Jones, D.A. Langs, C.M. Weeks, and H.A. Hauptman, Science **259**, 1430 (1993).

[5] C.M. Weeks, H.A. Hauptman, G.D. Smith, R.H. Blessing, M.M. Teeter, and R. Miller, Acta Crystallogr., Sect. D: Biol. Crystallogr. **D51**, 33 (1995).

[6] I. Abrahams, P. Lightfoot, and P.G. Bruce, J. Solid State Chem. **104**, 397 (1993).

[7] M. Louër, R. Brochu, D. Louër, S. Arsalane, and M. Ziyad, Acta Crystallogr., Sect. B: Struct. Sci. **B51**, 908 (1995).

[8] A. Clearfield, L.B. McCusker, and P.R. Rudolf, Inorg. Chem. **23**, 4679 (1984).

[9] J.P. Attfield, A.W. Sleight, and A.K. Cheetham, Nature (London) **322**, 620 (1986).

[10] D. Louër and M. Louër, J. Solid State Chem. **68**, 292 (1987).

[11] P. Lightfoot, A.K. Cheetham, and A.W. Sleight, Inorg. Chem. **25**, 3544 (1987); P. Lightfoot, C. Glidewell, and P.G. Bruce, J. Mater. Chem. **2**, 361 (1992).

[12] P. Benard, M. Louër, and D. Louër, J. Solid State Chem. **94**, 27 (1991).

[13] M. Tremayne, B.M. Kariuki, and K.D.M. Harris, Angew. Chem. **109**, 788 (1997); Angew. Chem. Int. Ed. Engl. **36**, 770 (1997).

[14] K. Shankland, W.I.F. David, and T. Csoka, Z. Kristallogr. **212**, 550 (1997).

[15] Y.G. Andreev, P. Lightfoot, and P.G. Bruce, J. Appl. Crystallogr. **30**, 294 (1997); Y.G. Andreev, G.S. MacGlashan, and P.G. Bruce, Phys. Rev. B **55**, 12 011 (1997); Y.G. Andreev and P.G. Bruce, J. Chem. Soc. Dalton Trans. **24**, 4071 (1998).

[16] W.P. Su, Acta Crystallogr., Sect. A: Found. Crystallogr. **A51**, 845 (1995); W.P. Su, Physica A **221**, 193 (1995); M.F. Zimmer and W.P. Su, Phys. Rev. E **58**, 5131 (1998); Y. Chen and W.P. Su, Acta Crystallogr., Sect. A: Found. Crystallogr. **A56**, 127 (2000).

[17] K.D.M. Harris and M. Tremayne, Chem. Mater. **8**, 2554 (1996); B.M. Kariuki, D.M.S. Zin, M. Tremayne, and K.D.M. Harris, *ibid.* **8**, 565 (1996); M. Tremayne, B.M. Kariuki, and K.D.M. Harris, J. Appl. Crystallogr. **29**, 211 (1996).

[18] R.E. Dinnebier, P.W. Stephens, J.K. Carter, A.N. Lommen, P.A. Heiney, A.R. McGhie, L. Brard, and A.B. Smith III, J. Appl. Crystallogr. **28**, 327 (1995).

[19] G. Reck, R.-G. Kretschmer, L. Kutschabsky, and W. Pritzkow, Acta Crystallogr., Sect. A: Found Crystallogr. **A44**, 417 (1988).

[20] J. Cirujeda, L.E. Ochando, J.M. Amigó, C. Rovira, J. Rius, and J. Veciana, Angew. Chem. ; **107**, 99 (1995); Angew. Chem. Int. Ed. Engl. **34**, 55 (1995).

[21] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, Science **220**, 671 (1983).

[22] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985); A.T. Brunger, J. Mol. Biol. **203**, 803 (1988).

[23] S.C. Lin and J.H.C. Hsueh, Physica A **203**, 369 (1994); **205**, 367 (1994).

[24] P. Slarry and G. Dreyfus, J. Phys. France **45**, L-39 (1983); S. R. White, in *Proceedings of the ICCD, 1984* (IEEE, New York, 1984), p. 646; S. Geman and D. Geman, IEEE, PAMI **6**, 721 (1984).

[25] A.V. Yatsenko, V.A. Tafeenko, K.A. Paseshnichenko, T.V. Stalnaya, and S.L. Solodar, Z. Kristallogr. **213**, 466 (1998).

[26] *International Tables for X-ray Crystallography. International Union of Crystallography*, edited by J. Ibers and W. C. Hamilton (The Kynoch Press, Birmingham, 1974).

[27] P. T. Beurskens, G. Admiraal, G. Beurskens, W. P. Bosman, R. de Gelder, R. Israél, and J. M. M. Smits, The DIRDIF Program System (Technical Report Crystallography Laboratory, University of Nijmegen, The Netherlands, 1994).

[28] G. M. Sheldrick, SHELXL-93, Program for the Refinement of Crystal Structures (University of Göttingen, Germany, 1993).

[29] W. Kraus and G. Nolze, computer code POWDERCELL for Windows, Ver. 2.1 (Federal Institute for Materials Research and Testing, Berlin, 1999).