PHYSICAL REVIEW E **73**, 015701(R) (2006)

Monte Carlo basin paving: An improved global optimization method

Lixin Zhan, Jeff Z. Y. Chen, and Wing-Ki Liu
Department of Physics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

(Received 12 August 2005; published 24 January 2006)

We propose a global optimization procedure, basin paving, which is based on the combination of the optimization strategies behind basin hopping and energy landscape paving. As an example, we describe its application in the protein structure prediction by examining two well-studied peptides, where we have found lower potential energy minima than previously located. We also compare the statistics of the searching trajectories produced by basin paving, basin hopping, and energy landscape paving.

DOI: 10.1103/PhysRevE.73.015701 PACS number(s): 87.15.−v, 02.60.Pn, 02.50.Ng, 05.10.Ln

The design of an effective global optimization algorithm, which performs a computational search for the global extremum of a multivariable target function, has been a longstanding problem in computational physics. Although rooted in more traditional physical problems, the development of more efficient methods, (where we are able to search through a vast number of local extrema that sometimes are isolated from each other in the variable space), has been continuously challenged by numerous topics of current interests in theoretical and applied physics, such as the prediction of protein structures [1–3], *ab initio* computation of nanosize atomic clusters [1], and optimization in transportation systems [4]. A Monte Carlo based optimization method is conceptually dependent on the introduction of hypothetical thermodynamics into the treatment, which converts the target function, if not already the potential energy, into an effective "potential energy" and allows thermal excitations to enable the search procedure to move around energy minima. Hence, we are dealing with the general problem of finding the global energy minimum in a complex energy landscape, where deep traps of local energy minima could seriously prevent the reaching of the global energy minimum within a reasonable computational time [1,5–10].

In this paper, we present an improved Monte Carlo global minimization method, basin paving, based on a combination of and the improvement over two state-of-the-art algorithms, basin hopping [5,6], and energy landscape paving [7]. Examining a few previously well-studied examples in the prediction of protein structures, we then demonstrate the efficiency of basin paving in browsing through the low-energy space, reflected by the discovery of the global energy minima.

We start by briefly reviewing the principle idea behind the *basin hopping* (BH) method [1,5,6] for minimization of a potential-energy function $E(\mathbf{r})$, where $\mathbf{r}$ is a multidimensional vector representing the generalized "atomic coordinates" of the system. The central scheme consists of two basic steps, iterated many times. Within the first step, a newly generated atomic configuration (based on small, random displacements of existing atomic coordinates) is used as the initial guess for the input of a typical *deterministic* minimization algorithm [11], that precisely locates its nearest local energy minimum. In the second step, this local energy minimum is then used as the current measure of energy for a typical Monte Carlo (MC) procedure, which evaluates the acceptability of the generated configuration based on a Boltzmann weight with a prespecified temperature. Essentially, BH maps the original energy landscape into a reduced energy landscape of a staircase form, where local energy minima appear as plateaus in a multidimensional space. This way, BH eliminates most of the high-energy barriers associated with the potential-energy maxima in the original energy function, while pinning down the precise values of energy minima at the same time. In fact, BH is a canonical (i.e., Boltzmann weight based) MC procedure applied on a reduced energy landscape transformed by a selected deterministic method. Despite the success of BH [1,5,6,12], when a system is complicated enough, the reduced energy landscape could still contain isolated energy traps, separated by high-energy plateaus; these energy traps hamper the computational efficiency and can be dealt with by further improving the second step mentioned above [9,10].

On the other hand, it has been realized a long time ago that an energy minimization procedure does not have to strictly follow a computational trajectory governed by the principle of detailed balance. The simulated thermodynamics, through the implementation of a statistical weight, in canonical [5,6] or other generalized ensembles [9,10], is a tool but not the final goal, for energy minimization. A directed search towards the desired global energy minimum can be encouraged by altering the weight function used in the search or equivalently by deforming the energy landscape to avoid regions that have already been explored. In an application of such an idea, Hansmann and Wille proposed the *energy landscape paving* (ELP) method [7], which utilized the core idea from a tabu search [13]. Instead of considering the minimization of $E(\mathbf{r})$, they considered the minimization of

$$\mathcal{E}(E(\mathbf{r}),t) \equiv E(\mathbf{r}) + f(E,t), \tag{1}$$

in which $f(E,t)$ is a paving function that depends on the history of a particular computational search and changes at each MC step. In practice, they have suggested using the accumulated histogram function from all previously visited energies at the MC step t, $H(E,t)$, to construct the paving function [7,14]. As a consequence, the searching process keeps track of the energies visited so that it can bias against revisiting those energies again in the immediate future. The additional paving function helps the simulation escape local entrapment and surpassing high-energy barriers easier. The paved energy is then used in the Boltzmann weight,

$$w(\mathbf{r},t) = \exp\{-\mathcal{E}(E(\mathbf{r}),t)/k_BT\}, \tag{2}$$

for a typical MC search. Note that thermal fluctuations depend on the simulation temperature T, which cannot be exactly zero; hence the procedure yields only the approximate locations of energy minuma within an error of magnitude k_BT, where k_B is the Boltzmann constant.

Our *basin paving* (BP) method is based on a combination of BH and ELP, with a critical revision to the latter. After random initialization of the atomic coordinates, the computational procedure mainly includes the following steps, repeated iteratively. The first step is identical to that in BH. Based on the existing atomic coordinates computed from the previous iteration, $\mathbf{r}_{\min}$, we consider an attempted move by displacing the atoms from $\mathbf{r}_{\min}$ to a new configuration $\mathbf{r}'$. The coordinates $\mathbf{r}'$ are then used as the initial guess in a deterministic minimization algorithm to locate the nearest local minimum at $\mathbf{r}'_{\min}$. Under this step, the mathematical equivalence is the transformation of the original energy landscape $E(\mathbf{r})$ into $\widetilde{E}(\mathbf{r})$, which contains only the local minima of $E(\mathbf{r})$,

$$\widetilde{E}(\mathbf{r}) \equiv \min\{E(\mathbf{r})\} = E(\mathbf{r}_{\min}). \tag{3}$$

In the next step, we consider an axillary energy function,

$$\widetilde{\mathcal{E}}(\mathbf{r},t) \equiv \widetilde{E}(\mathbf{r}) + \epsilon H(\widetilde{E},t), \tag{4}$$

where ϵ is an adjustable constant. The acceptability of the new configuration $\mathbf{r}'$ is determined by a comparison between $\widetilde{E}(\mathbf{r}')$ and $\widetilde{E}(\mathbf{r})$, where two cases are possible: (a) $\widetilde{E}(\mathbf{r}') < \widetilde{E}(\mathbf{r})$ and (b) $\widetilde{E}(\mathbf{r}') \geqslant \widetilde{E}(\mathbf{r})$. For case (a), we unconditionally accept the new configuration and return to the first step to start a new iteration; for case (b), we consider a transition probability of

$$P = \exp\{[\widetilde{\mathcal{E}}(\mathbf{r},t) - \widetilde{\mathcal{E}}(\mathbf{r}',t)]/k_BT\} \tag{5}$$

to decide whether the attempted move is acceptable.

This procedure contains the unconditional acceptance of the new configuration in case (a), which can be contrasted with the original ELP where the transition probability of Eq. (5) is *always* used for both (a) and (b). It turns out that this revision has a profound consequence in the efficiency of the algorithm as will be discussed below by examples. Conceptually, the revision is targeted at overcoming a technical flaw in ELP—the collection of the histogram function $H(E,t)$ is actually performed by dividing the energy space into finite-size bins. Consider an attempted move in ELP which yields a new, lower-energy minimum that has never been visited before and happens to fall into the same bin containing other energies previously visited in earlier steps. Undesirably, the likelihood of accepting this new energy minimum becomes small if the histogram function corresponding to that specific bin is large. Our newly revised, unconditional acceptance for case (a) overcomes this pitfall. As examples below show, our revision meets the requirement of extensively searching the low-energy space and yet not overly exaggerating the sampling of the low-energy region.

A native conformation of a protein is often believed to be close to its lowest potential-energy configuration [15]. At present, because of the limitation in computational power, quantum-mechanical calculations of the potential-energy surface of such systems by solving the many-body Schrödinger equation is impractical, even for a short peptide. Instead, a common approach is to represent the potential energy of a system by a classical force field, in which parameters are empirically determined to best fit experimental results and *ab initio* calculations for small systems. To determine the native structure, one performs a global optimization to find the global minimum of the potential energy as a function of the coordinates of the representative atoms for a given amino acid sequence.

To benchmark the effectiveness of BP, we determine the energy minima of two well-studied peptides, the pentapeptide Met-enkephalin, which has only 5 residues [6,7,16–19], and the villin headpiece subdomain HP-36, a relatively long sequence with 36 residues [7,20–22]. In this study, we separately used both versions of the empirical confirmational energy program for peptides (ECEPP) force field, ECEPP/2 and ECEPP/3 [23] (which differ in parametrization), for the descriptions of the potential energy between the representative atoms in the sequences. Typically, ECEPP contains two-body potentials, representing electrostatic, hydrophobic, and hydrogen-bonding interactions in the system. The relative positions of atoms are entirely determined by the torsional angles between atomic bonds in the system. The bondlengths and bond angles are fixed at experimental values and out-of-plane deformation of peptide bonds is not permitted. The software package SMMP [24] was adopted in this work for computing the ECEPP potentials.

In our first example, we consider the pentapeptide Met-enkephalin capped by NH_2 for the N terminus and COOH for the C terminus. Computationally, this peptide is one of the most frequently studied examples, and the lowest energies have been reported previously [6,16–18]. In total, we considered four versions of ECEPP: $E/2_\pi$, $E/2$, $E/3_\pi$ and $E/3$, where $E/2$ and $E/3$ are abbreviates for ECEPP/2 and ECEPP/3, respectively, and a subscript π is an indication that the backbone dihedral angles ω are constrained to π. The lowest energies found by our BP search and their corresponding conformations, for these four cases, are shown in Figs. 1(b)–1(e). Their dihedral angles can be found in Ref. [25]. Previously, the local minima with energies not much higher than the global minimum for case $E/2$ were sampled and classified by Freyberg and Braun [16], using BH with a variable temperature parameter. The lowest energy was found to be -12.91 kcal/mol [16] which was reproduced by our calculation reported in Fig. 1(c). For cases with ω fixed, Eisenmenger and Hansmann [17] have reported their finding of the lowest-energy minima of -10.72 kcal/mol [17] and -10.85 kcal/mol [17] for $E/2_\pi$ and $E/3_\pi$, respectively, using a multicanonical method. While our result in Fig. 1(b) confirms theirs for $E/2_\pi$, a new energy minimum $E = -10.90$ kcal/mol, which is about 0.05 kcal/mol lower than that found in Ref. [17], has been obtained in this work for $E/3_\pi$. The significance of our result goes beyond a merely new lower energy—the value of this energy, no matter how precisely determined, is an approximation to the true minimum energy since approximations are introduced in proposing the force field. More important than the energy differ-

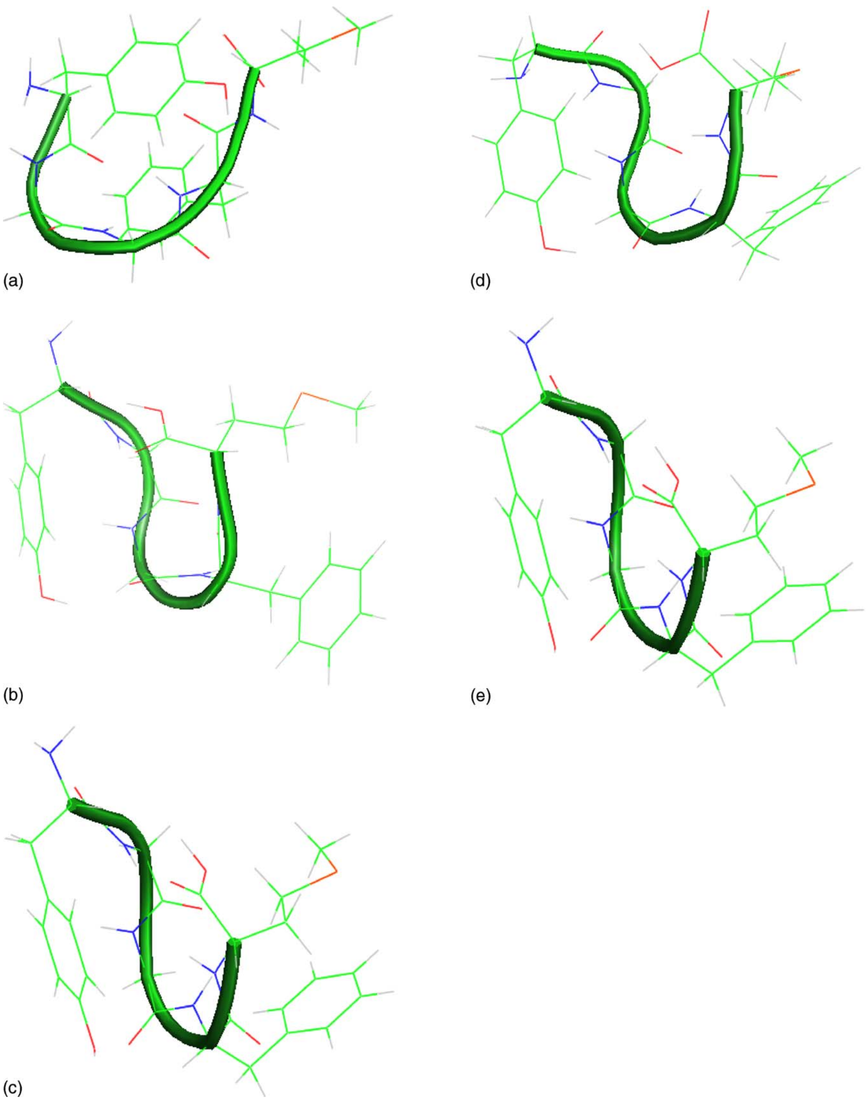

FIG. 1. (Color online) Comparison between the predicted lowest energy structures of Met-enkephalin, visualized by a backbone tube plot. Eisenmenger and Hansmann [17] have determined the structure shown in (a) using the multicanonical method, while we obtained the structure in (d) with a lower-energy minimum using BP, both based on the same version of ECEPP, $E/3_{\pi}$ (see text). In (b), (c), and (e) we show the structures determined using BP based on $E/2_{\pi}$, $E/2$, and $E/3$, the first two verifying the results from Refs. [16,17]. The energies (in kcal/mol) associated with the structures are: (a) −10.85, (b) −10.72, (c) −12.91, (d) −10.90, and (e) −12.43.

ence, there exists a distinctive structural difference between the configurations, which can be visually inspected between Figs. 1(a) and 1(d). Furthermore, by comparing the structures in Figs. 1(b)–1(e), we found an overall consistency in the prediction of the native structure of Met-enkephalin, based on various versions of ECEPP.

The ability of BP in capturing new energy minimum, exemplified above, lies in its more thorough search in the low-energy regime. To demonstrate this and to consider the possible influence of the parameter k_BT used in the algorithm [see Eq. (5)], we plotted the histograms of energies visited for case $E/3_{\pi}$ by BP over 2×10^4 MC steps for Met-enkephalin at T=5, 50, 500, 1 000, and 2 000 K, with ϵ =1 kcal/mol. We found a few features in BP that have more advantages over other methods. First, despite the rather large temperature range, the solid curves in Fig. 2 display a fairly consistent general shape for the energy histograms. This is because that in BP, apart from the very initial searching period, the histogram difference, rather than the energy difference, dominates the selection rule most of the time for moving uphill. This can be contrasted with other MC based simulation techniques where temperature is a vital parameter. Secondly, BP extensively focuses the search on the low-energy region but retains the ability of sweeping the high-energy region to overcome the energy barriers. For comparison, a typical energy histogram of BH has a peak at a mean energy determined by k_BT (see the dashed curve in Fig. 2), which is known to be inefficient in the low-energy region. As

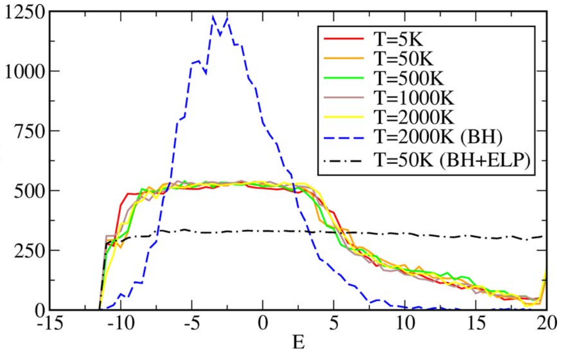

FIG. 2. (Color online) Histograms of energies searched by BP at various temperatures (solid curves), by BH at T=2 000 K (dashed), and by ELP on the reduced energy landscape at T=50 K (dot-dashed), within the first 2×10^4 MC steps, for Met-enkephalin, based on $E/3_{\pi}$.

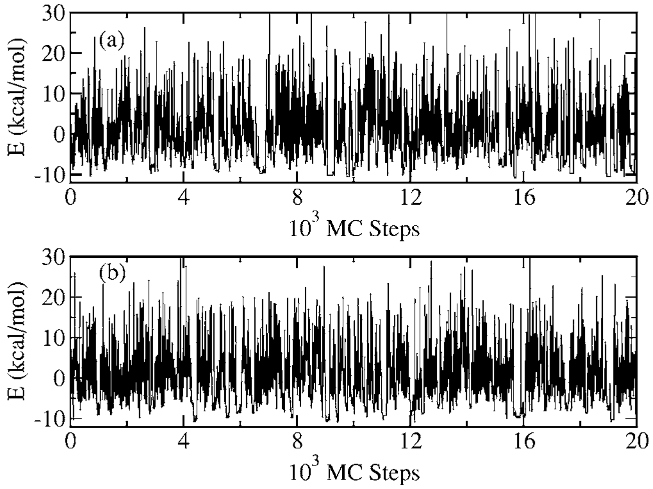

FIG. 3. Typical BP search trajectories at (a) $T=5$ K and (b) $T=2\,000$ K.

another comparison, we have also plotted the energy histogram (dot-dashed curve in Fig. 2) from an ELP run based on the reduced energy landscape; the transition probability P in Eq. (5) was used for both cases, (a) and (b), discussed above Eq. (5). The unconditional acceptance of case (a), when a lower energy is produced in comparison with that in the previous iteration step, is our key revision to ELP that encourages the return to the low-energy region. Finally, a further illustration of the BP search process is given in Fig. 3, where we plot the visited energy at every MC step. We can see from these trajectories that BP is able to frequently switch between the search regions, in high and low energies, with almost no dependence on the temperature parameter used.

To test BP in a larger system, we consider as our second example villin headpiece subdomain HP-36 containing 587 representative atoms of the 36 residues capped by C and N terminuses. Previously many attempts have been made to study the folding of this peptide numerically [7,21,22]. Using the parallel tempering method and the ECEPP/2 force field, Lin *et al.* has determined an energy minimum of $E=-209.2$ kcal/mol [22]. For comparison, based on a fully stretched initial conformation, we have used BP to compute the energy minimum based on the same force field. In our simulations, we are able to locate a lower-energy minimum, $E=-209.65$ kcal/mol, obtained in one of the 10 BP runs consisting of 2×10^5 MC trials. Also yielding other low-energy minima, the computation located a structure similar to that determined by an NMR experiment [20]. Details on the structure determination will be reported elsewhere.

To further illustrate the improvement in efficiency, BP was applied to Lennard-Jones clusters containing N atoms. The mean MC steps based on 10 independent BP runs to locate the global minima are 10 851 for $N=150$, and 34 361 for $N=155$. In contrast, it takes about twice the MC steps to locate the global minima for the same systems based on the BH method [9].

In summary, we have proposed a new Monte Carlo optimization method based on the ideas behind basin hopping [5,6] and energy landscape paving [7]. With a critical revision to the latter, our method has the ability of surpassing high-energy barriers and searching the lower-energy region in more detail. The success of the method can be shown by its ability of finding new energy minima, lower than previous results, of two of the best known examples in protein structure computation.

The authors thank NSERC for providing financial support and SHARCNET for allocating computational time.

[1] D. J. Wales and H. A. Scheraga, Science **285**, 1368 (1999).
[2] D. Baker, Nature (London) **405**, 39 (2000).
[3] D. Baker and A. Sali, Science **294**, 93 (2001).
[4] *Mathematical Methods on Optimization in Transportation Systems*, edited by M. Pursula and J. Niittymaki (Kluwer Academic Publishers, Dordrecht, 2001).
[5] D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111 (1997).
[6] Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
[7] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).
[8] S. Kirkpatrick *et al.*, Science **220**, 671 (1983); B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977); J. Lee, Phys. Rev. Lett. **71**, 211 (1993); **71**, 2353(E) (1993); R. H. Swendsen and J.-S. Wang, *ibid.* **57**, 2607 (1986); P. M. C. de Oliveira *et al.*, Braz. J. Phys. **26**, 677 (1996); J.-S. Wang and L. W. Lee, Comput. Phys. Commun. **127**, 131 (2000); F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001); S. Jang *et al.*, *ibid.* **91**, 058305 (2003).
[9] L. Zhan *et al.*, J. Chem. Phys. **120**, 5536 (2004).
[10] L. Zhan *et al.*, J. Chem. Phys. **122**, 244707 (2005).
[11] P. E. Gill *et al.*, *Practical Optimization* (Academic Press, London, 1981).
[12] R. H. Leary and J. P. K. Doye, Phys. Rev. E **60**, R6320 (1999).
[13] F. Glover, ORSA J. Comput. **1**, 190 (1989); **2**, 4 (1990); D. Cvijovic and J. Klinowski, Science **267**, 664 (1995).
[14] A. Schug *et al.*, J. Chem. Phys. **122**, 194711 (2005).
[15] C. B. Anfinsen, Science **181**, 223 (1973).
[16] B. von Freyberg and W. Braun, J. Comput. Phys. **12**, 1065 (1991).
[17] F. Eisenmenger and U. H. E. Hansmann, J. Phys. Chem. B **101**, 3304 (1997).
[18] I. P. Androulakis *et al.*, J. Global Optim. **11**, 1 (1997).
[19] D. A. Evans and D. J. Wales, J. Chem. Phys. **119**, 9947 (2003).
[20] C. J. McKnight *et al.*, J. Mol. Biol. **260**, 126 (1996).
[21] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[22] C.-Y. Lin *et al.*, Proteins **52**, 436 (2003).
[23] G. Nemethy *et al.*, J. Phys. Chem. **87**, 1883 (1983); G. Nemethy *et al.*, *ibid.* **96**, 6472 (1992).
[24] F. Eisenmenger *et al.*, Comput. Phys. Commun. **138**, 192 (2001).
[25] L. Zhan, Ph.D. thesis, University of Waterloo, 2005 (unpublished).