# Conformational rigidity in a lattice model of proteins

Olivier Collet
*Equipe de Dynamique des Assemblages Membranaires, UMR CNRS 7565, Faculté des Sciences, Université Henri Poincaré-Nancy 1, 54506 Vandoeuvre-lès-Nancy, France*

(Received 22 January 2003; published 26 June 2003)

It is shown in this paper that some simulations of protein folding in lattice models, which use an incorrect implementation of the Monte Carlo algorithm, do not converge towards thermal equilibrium. I developed a rigorous treatment for protein folding simulation on a lattice model relying on the introduction of a parameter standing for the rigidity of the conformations. Its properties are discussed and its role during the folding process is elucidated. The calculation of thermal properties of small chains living on a two-dimensional lattice is performed and a Bortz-Kalos-Lebowitz scheme is implemented in the presented method in order to study kinetics of chains at very low temperature. The coefficients of the Arrhenius law obtained with this algorithm are found to be in excellent agreement with the value of the main potential barrier of the system. Finally, a scenario of the mechanisms, including the rigidity parameters, that guide a protein towards its native structure, at medium temperature, is given.

DOI: 10.1103/PhysRevE.67.061912 PACS number(s): 87.14.Ee, 87.15.He, 87.15.Aa, 05.10.Ln

## I. INTRODUCTION

Proteins are heteropolymers that exhibit surprising thermodynamic and kinetic properties: the conformation of lowest free energy which is assumed to be a unique, stable, and biologically active structure [1] is found in very short times. A major challenge in theoretical protein folding is to understand the kinetic aspect, i.e., under physiological conditions, how does a protein find its native structure in biologically reasonable times [2]. Simulations using full atomistic representation of the protein and the solvent coupled to a molecular dynamic algorithm have been widely used to study this problem [3,4]. But, due to the large number of water molecules around a protein and the sophistication of the force fields used to calculate the energy of the system, such approaches are very time consuming. To sample, more widely, the conformational space, it is more efficient to ‘‘preaverage’’ the solvent and treat it implicitly by adding solvation terms to the potential energy of the heteropeptide [5–11]. But, even with such solvation models, calculations of the partition function of a protein still remain illusive.

It is then necessary to reduce the representation of the problem further and the lattice model is a class of coarse-grained models that is often used to study theoretically the folding of a protein [12–20]. In such approaches, the amino acids of the chain are positioned on a square [12,21] or cubic [15] grid and the intrachain energy is the summation of all the pairwise contributions between residues. Two main different models of potential have been widely used in simulations: the $HP$ energy model in which a monomer is either hydrophobic ($H$) or polar ($P$) [12,22] or the random energy model (REM) [15,23,24]. For a not too large chain, the conformation of minimum of energy can be easily found by exhaustive enumeration [25] and for a large chain the native structure can be derived from selected sequences [24]. Chan and Dill [19,26] simulated the folding of a protein on a lattice using the $HP$ model where the evolution of the probability of occurrence of each conformation is obtained by performing products of the matrix of connections between conformations. In that work, they had introduced one normalization constant applied to all conformations to guarantee detailed balance. However, the convergence towards thermal equilibrium of such approaches has never been checked by long simulations and the physical meaning of the normalization constant has never been questioned.

On the other hand, most of the kinetic studies of the protein folding on lattice models have been performed using Monte Carlo (MC) algorithms [27–29] applied to the $HP$ model of energy [22,30,19,26] or REM [31,32,17,33–37]. Very few papers describe in detail the algorithm used to generate the trial moves in lattice models. In some of them, the MC method applied to lattice polymers does not obey detailed balance conditions because it used a nonsymmetric probability matrix to generate the trial moves. Doubts about these procedures have been raised by Sorenson and Head-Gordon [38] and by Kaya and Chan [20].

In this paper, it is shown that some implementations of the MC algorithm for lattice models violate the detailed balance conditions and that such simulations do not converge towards thermal equilibrium. An attempt to refine the algorithm has been recently proposed [39]. This method converges towards equilibrium, but the parameters found for the Arrhenius law disagree with the value of the main potential barrier obtained independently by a study of the phase space of the system. The purpose of this work is not to find a new implementation of MC method which gives shorter or longer folding times than those obtained with other algorithm, but rather to solve a problem of convergence of the simulation towards thermal equilibrium with a correct probability distribution of the conformations. A rigorous treatment of the dynamics which leads to an efficient sampling of the conformational space, a precise calculation of kinetic parameters and the determination of the correct Arrhenius law has been introduced. The introduction of a parameter depending on the conformations, based on a rewriting of the detailed balance condition, in the algorithm implied a good convergence towards thermal equilibrium. Moreover, the mechanism that

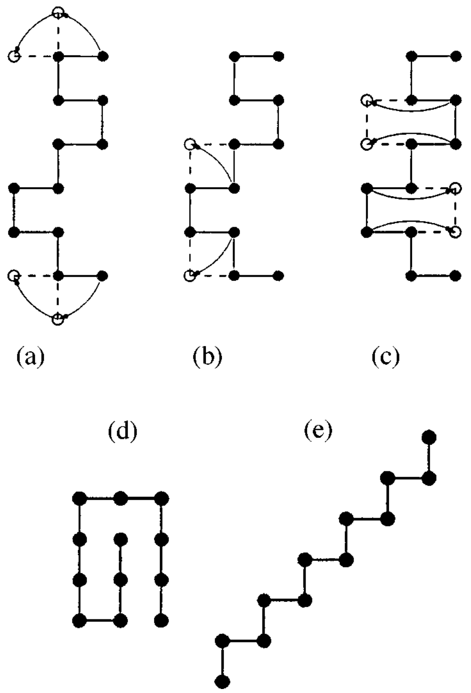

FIG. 1. Types of moves allowed by the algorithm: (a) tail move, (b) corner flip, (c) crankshaft move. The solid lines and the full circles are for the chain and the monomers. The dashed lines and the empty circles are for the bonds and the monomers of the chain affected by a move. The conformation with (d) the smallest number (only 1) of neighbors and (e) the largest number $(N-2)$ of neighbors.

guides a chain towards its native structure at medium temperature is also discussed.

## II. MODEL AND METHOD

The model used in this work is a two-dimensional lattice polymer. The self-avoiding chains composed of $N$ monomers are constrained to be on a square lattice. The energy of a given conformation $m$ is given by

$$E^{(m)}=\sum_{i>j+1}(B_{ij}+B_0)\Delta_{ij}^{(m)}, \tag{1}$$

where the function $\Delta_{ij}^{(m)}=1$, if the $i$th and $j$th monomers interact, i.e., if they are first neighbors on the lattice, and $\Delta_{ij}^{(m)}=0$ otherwise. The $B_{ij}$'s are the contact energy values chosen randomly in a Gaussian distribution centered on 0 and give the sequence of the chain. The parameter $B_0$ is chosen equal to $-1$ to favor the compact conformations [15,21].

In the MC simulations used in this paper, the sets of connections between conformations are those defined by Chan and Dill [19], where the rigid rotation chain, given in Fig. 1b(ii) of Ref. [19], has been removed. The tail move [see Fig. 1(a)] and the corner flip [see Fig. 1(b)] are referred to as the move set $a$ ($\mathrm{MS}_a$), the crankshaft move [see Fig. 1(c)] is referred to as the move set $b$ ($\mathrm{MS}_b$). The evolution of the chain is carried out by performing local modifications of the conformations using $\mathrm{MS}_a$ and $\mathrm{MS}_b$.

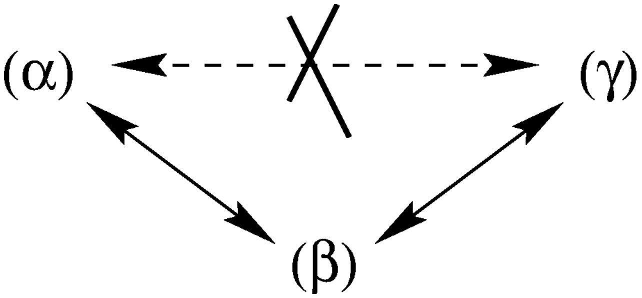

FIG. 2. Hypothetical system where the number of connections (solid arrows) depends on the conformations, because there is no connection between $\alpha$ and $\gamma$. An example of a typical trajectory at very high temperature can be $\alpha\beta\alpha\beta\gamma\beta\alpha\beta\gamma\beta\gamma\beta\gamma\beta\alpha\beta\gamma\beta\cdots$, where $\beta$ occurs one step over two.

As can be seen in the example shown in Fig. 1, whereas only one move is allowed for conformation (d), 14 moves are allowed for conformation (e), showing that using such move sets the number of allowed connections depends on the conformations. This particularity of lattice models induces nonconvergence towards equilibrium in a simulation using a MC algorithm, in which a modification is proposed at each step and is performed following a criterion of acceptance (as the test of Metropolis, for example). An illustration of this point is given by the very simple three-state system shown in Fig. 2. At very high temperature, the equilibrium probabilities of the three conformations must be equal and in a simulation the occurrence of each conformation should equal 1/3. However, as any proposed move is always accepted by the Metropolis test at very high temperature, a typical random trajectory gives a probability of occurrence of 1/2 for the conformation $\beta$ and of 1/4 for each of conformations $\alpha$ and $\gamma$. This problem arises from the fact that the number of connections can be different for each conformation.

The purpose of this work is to propose a correct MC simulation [27] for a lattice model using $\mathrm{MS}_a$ and $\mathrm{MS}_b$, which guarantees the convergence towards thermal equilibrium imposed by the condition of the detailed balance [29]:

$$P_{\mathrm{eq}}^{(m)}W(m\rightarrow n)=P_{\mathrm{eq}}^{(n)}W(n\rightarrow m), \tag{2}$$

where

$$P_{\mathrm{eq}}^{(m)}\propto\exp(-E^{(m)}/T) \tag{3}$$

is the equilibrium probability of the conformation $m$ and $T$ is the temperature. The transition probability from the state $m$ to the state $n$ can be rewritten:

$$W(m\rightarrow n)=W^{(0)}(m\rightarrow n)a(m\rightarrow n), \tag{4}$$

where $W^{(0)}(m\rightarrow n)$ is the *a priori* transition probability, i.e., the probability to select the move $m\rightarrow n$ and $a(m\rightarrow n)$ is the acceptance rate of the transition $m\rightarrow n$ which indicates if it is performed or not. Then, with the convenient choice for the acceptance ratio,

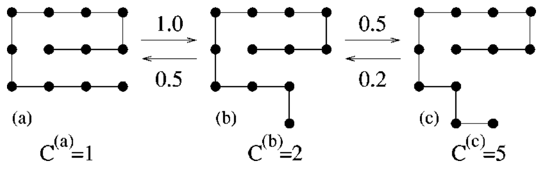

FIG. 3. A part of the connection graph of the 12 monomer chain. In $\mathrm{MC}_0$ implementation where the *a priori* transition probabilities $W_0(m \rightarrow n)$ (shown above the arrows) are not symmetric because they depend on the number of neighbors. Such an algorithm gives a ratio of the occurrence of the conformation (b) over that of (a) two times larger than the ratio of their equilibrium probabilities.

$$a(m \rightarrow n) = \frac{1}{1+\exp(\Delta E_n^m/T)}, \tag{5}$$

where $\Delta E_n^m = E^{(n)} - E^{(m)}$, the *a priori* transition probabilities must have a symmetric form,

$$W^{(0)}(m \rightarrow n) = W^{(0)}(n \rightarrow m) \tag{6}$$

to satisfy Eqs. (2)–(5).

The total number of allowed transitions from $m$ by performing a move of $\mathrm{MS}_a$ ($\mathrm{MS}_b$) is noted $C_a^{(m)}$ ($C_b^{(m)}$). Some simulations previously proposed [17], noted in the following as $\mathrm{MC}_0$, violates this condition. In these simulations, at each step of a $\mathrm{MC}_0$ implementation, a move of $\mathrm{MS}_a$ ($\mathrm{MS}_b$) is always selected among the $C_a^{(m)}$ ($C_b^{(m)}$) possible ones of the current conformation $m$ with a probability $r$ $(1-r)$. The *a priori* probabilities to select the transition $m \rightarrow n$ with $\mathrm{MS}_a$ and $\mathrm{MS}_b$ are given by the following two equations:

$$W_a^{(0)}(m \rightarrow n) = r\,\frac{\delta_a^{mn}}{C_a^{(m)}}, \tag{7}$$

$$W_b^{(0)}(m \rightarrow n) = (1-r)\,\frac{\delta_b^{mn}}{C_b^{(m)}}, \tag{8}$$

where $\delta_a^{mn}=1$ ($\delta_b^{mn}=1$) if the conformations $m$ and $n$ are connected by a move of $\mathrm{MS}_a$ ($\mathrm{MS}_b$) and $\delta_a^{mn}=0$ ($\delta_b^{mn}=0$) otherwise. As the quantities $W_a^{(0)}(m \rightarrow n)$ and $W_b^{(0)}(m \rightarrow n)$ depend on the number of connections, $C_a^{(m)}$ and $C_b^{(m)}$, of conformation $m$, one obtains

$$W_a^{(0)}(m \rightarrow n) \neq W_a^{(0)}(n \rightarrow m) \quad \text{if} \quad C_a^{(m)} \neq C_a^{(n)}, \tag{9}$$

$$W_b^{(0)}(m \rightarrow n) \neq W_b^{(0)}(n \rightarrow m) \quad \text{if} \quad C_b^{(m)} \neq C_b^{(n)}. \tag{10}$$

Figure 3 shows a part of a connection graph where the probabilities, $W_a^{(0)}(m \rightarrow n)$ and $W_b^{(0)}(m \rightarrow n)$, depend on the conformation $m$.

To solve this problem, a normalization of *a priori* probability of transition is introduced. The two conformations with the largest number of neighbors using either $\mathrm{MS}_a$ or $\mathrm{MS}_b$ are shown in Fig. 4 and the maximum number of moves allowed by $\mathrm{MS}_a$ or $\mathrm{MS}_b$ (related to these structures) are

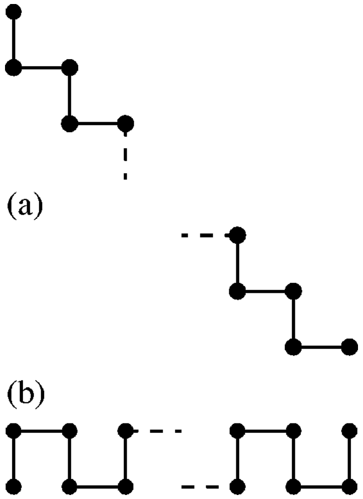

FIG. 4. (a) The conformation with the maximum number of connections allowed by $\mathrm{MS}_a$. (b) The conformation with the maximum number of connections allowed by $\mathrm{MS}_b$.

$$C_a^{\max} = \max_m \{C_a^{(m)}\} = N+2, \tag{11}$$

$$C_b^{\max} = \max_m \{C_b^{(m)}\} = N-7. \tag{12}$$

In the implementation, noted MC*, proposed in this work, the *a priori* probabilities to attempt a move from conformation $m$ to conformation $n$ are rewritten:

$$W_a^{(0)}(m \rightarrow n) = \frac{r}{C_a^{\max}}\,\delta_a^{mn} = \frac{r}{N+2}\,\delta_a^{mn}, \tag{13}$$

$$W_b^{(0)}(m \rightarrow n) = \frac{(1-r)}{C_b^{\max}}\,\delta_b^{mn} = \frac{1-r}{N-7}\,\delta_b^{mn} \tag{14}$$

and do not depend on the conformations. To fix the value of $r$, it is assumed that, in contrast to a rigid rotation [19] that involves movement of a lot of monomers, the one and two monomer moves are local modifications having then the same affinity. It follows from Eqs. (13) and (14) that $r/(N+2) = (1-r)/(N-7)$, leading to

$$r = \frac{N+2}{2N-5}. \tag{15}$$

The transition probability becomes

$$W^{(0)}(m \rightarrow n) = W_a^{(0)}(m \rightarrow n) + W_b^{(0)}(m \rightarrow n) \tag{16}$$

$$= \frac{\delta^{mn}}{2N-5} \quad \text{with} \quad \delta^{mn} = \delta_a^{mn} + \delta_b^{mn}. \tag{17}$$

These quantities are well symmetric and this is the main difference with $\mathrm{MC}_0$ implementation, as will be discussed at the end of this section.

The *a priori* probability to attempt any move from the conformation $m$ is then

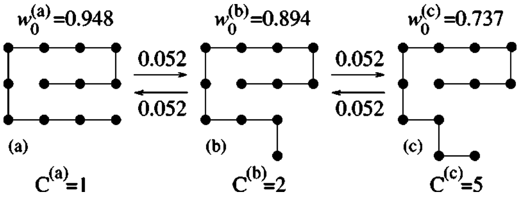

FIG. 5. A part of the connection graph of the 12 monomer chain. The conformations (a), (b), and (c) are connected to, respectively, one, two, and five (not shown) neighbors, by $\mathrm{MS}_a$ (see text). In the proposed method, the probability to attempt a move, $W_0(m\to n) = 1/(2N-5) = 0.052$, is symmetric. The probability to not attempt a move during one MC* step, $w_0^{(m)}$, depends on the conformation.

$$\sum_{n\neq m} W^{(0)}(m\to n)=\frac{C^{(m)}}{2N-5}\neq 0 \quad \text{with} \quad C^{(m)}=C_a^{(m)}+C_b^{(m)}. \tag{18}$$

This quantity is never equal to 1 and depends on the conformation $m$. Therefore, there appears a probability of null transition,

$$w_m^{(0)}=1-\frac{C^{(m)}}{2N-5}\neq 0, \tag{19}$$

which is the *a priori* probability to not attempt a move from the conformation $m$ during one MC* step. The same part of the connection graph shown on Fig. 3 is shown on Fig. 5, including now the rigidity of each conformation. As these factors only depend on the conformation of the chain, they are sequence independent. To give a physical meaning of these parameters, one must note that the larger the parameter $w_m^{(0)}$ is, the more MC* steps are spent without attempting a move, then $w_m^{(0)}$ can be viewed as the rigidity of the conformation $m$.

In contrast to $\mathrm{MC}_0$ implementation, a MC* step of the proposed method consists in first choosing if a move is tried or not and, second, if a move is tried, selecting if this move is performed or not.

## III. RESULTS

In order to check the accuracy of the MC* in a reasonable computational time, it has been applied to a 12 monomer chain. This chain can adopt 15 037 different self-avoiding walk conformations nonequivalent by symmetry. Results shown in this paper are obtained for the sequence presented in Table I for both methods. Because the acceptance ratio for any given connection is the same in both methods, the transition probability of a transition is always smaller in MC* simulations. Then, a larger number of Monte Carlo steps is necessary to generate an accepted move using MC* rather than using $\mathrm{MC}_0$. As a result, to perform a given number of accepted moves, the total number of MC steps must be larger with MC*. The CPU cost of this first test is negligible compared to the Metropolis one. Then the CPU time used by these two methods is almost equal.

MC trajectories of 30 billion steps have been performed. For some given temperatures, convergence factor

$$C(t)=\sqrt{\sum_m [P_{\mathrm{eq}}^{(m)}-\pi^{(m)}(t)]^2}$$

is computed each 100 000 MC steps $t$ is the number of MC steps and $\pi^{(m)}(t)=n^{(m)}(t)/t$, where $n^{(m)}(t)$ is the number of MC steps for which the conformation $m$ occurs. $\pi^{(m)}(t)$ is the mean occurrence of conformation $m$ during the first $t$ steps of the MC simulations. If an algorithm fulfills the detailed balance, $C(t)$ should tend towards 0 when $t\to\infty$.

Results obtained with $\mathrm{MC}_0$ are first discussed. As was seen in Fig. 3, the probability that a conformation occurs at equilibrium in $\mathrm{MC}_0$ simulation is proportional to its number of connections times its equilibrium probability. The values of the mean occurrence of conformation $m$ for very large $t$ is then

$$\pi_\infty^{(m)}\propto P_{\mathrm{eq}}^{(m)}(1-w_m^{(0)}) \quad \text{for} \quad t\to\infty$$

and, after normalization of the values of $\pi_\infty^{(m)}$, the theoretical limit of the convergence factor is

TABLE I. The $B_{ij}$ couplings of the Gaussian sequence used in this paper.

| $B_{ij}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 0.0 | 0.0 | −0.631 | 0.0 | −20.047 | 0.0 | −0.750 | 0.0 | −1.321 | 0.0 | −0.529 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | −2.383 | 0.0 | −1.492 | 0.0 | −0.159 | 0.0 | −1.207 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | −1.171 | 0.0 | 0.122 | 0.0 | −0.900 | 0.0 | −0.461 |
| 4 | −0.631 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | −0.458 | 0.0 | −1.963 | 0.0 | −1.598 | 0.0 |
| 5 | 0.0 | −2.383 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | −1.568 | 0.0 | −0.880 | 0.0 | −0.990 |
| 6 | −20.047 | 0.0 | −1.171 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.205 | 0.0 | −1.208 | 0.0 |
| 7 | 0.0 | −1.492 | 0.0 | −0.458 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | −0.381 | 0.0 | −1.892 |
| 8 | −0.750 | 0.0 | 0.122 | 0.0 | −1.568 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | −1.650 | 0.0 |
| 9 | 0.0 | −0.159 | 0.0 | −1.963 | 0.0 | 0.205 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | −0.099 |
| 10 | −1.321 | 0.0 | −0.900 | 0.0 | −0.880 | 0.0 | −0.381 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 11 | 0.0 | −1.207 | 0.0 | −1.598 | 0.0 | −1.208 | 0.0 | −1.650 | 0.0 | 0.0 | 0.0 | 0.0 |
| 12 | −0.529 | 0.0 | −0.461 | 0.0 | −0.990 | 0.0 | −1.892 | 0.0 | −0.099 | 0.0 | 0.0 | 0.0 |

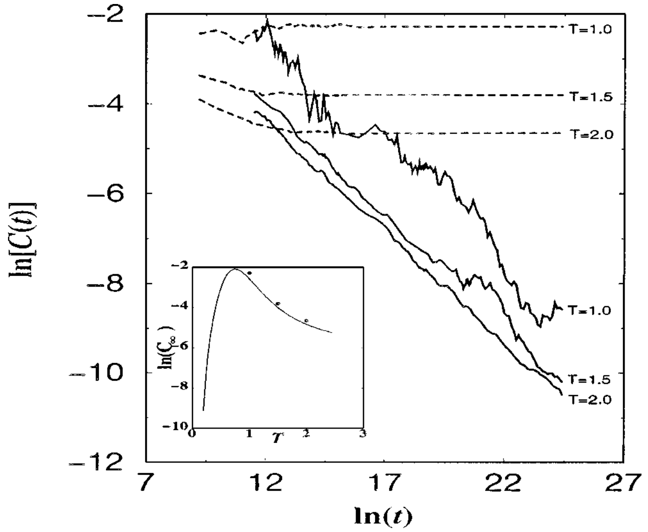

FIG. 6. Main plot: log-log plots of the convergence factor $C(t)$ versus the number of MC steps $t$ for different temperatures. Dashed lines: with $\mathrm{MC}_0$ for which the $W^{(0)}(m\rightarrow n)$ prefactor and $w_m^{(0)}$ parameter are omitted. Solid lines: with MC*. Inset: solid line, logarithm of the theoretical value of the limit of the convergence factor versus temperature [computed using Eq. (11)] for the commonly used method. The dots are the numerical values obtained by simulations at different temperatures using this method (limit values of the dashed lines of the main graph).

$$C(t)\underset{t\rightarrow\infty}{\rightarrow} C_\infty(T)=\sqrt{\sum_m\left(P_{\mathrm{eq}}^{(m)}-\frac{P_{\mathrm{eq}}^{(m)}(1-w_m^{(0)})}{\sum_{m'}P_{\mathrm{eq}}^{(m')}(1-w_{m'}^{(0)})}\right)^2}$$

$$\neq 0. \tag{20}$$

Figure 6 shows that the $\mathrm{MC}_0$ present clearly the limits of convergence depending on the temperature. The curve of $C_\infty(T)$ calculated with Eq. (20) and the numerical results obtained for $C(t)$ with large values of $t$ are in very good agreement (see Fig. 6, inset). This result shows that the probabilities of occurrence of conformations do not converge towards equilibrium probability computed using their Boltzmann weights. As a result, $\mathrm{MC}_0$ do not converge towards thermal equilibrium, and then, cannot be used efficiently to calculate thermodynamic or kinetic properties.

On the same figure, MC* shows a power law convergence: $C(t)\propto t^{-1/2}$ and then $C(t)\rightarrow 0$ for $t\rightarrow\infty$. This result shows clearly that the factors of null transition $w_m^{(0)}$ cannot be omitted in lattice simulations. They guarantee a correct sampling of the conformational space and the convergence of the simulation towards thermal equilibrium.

The accuracy of MC* for kinetic studies is now considered. A major problem in protein folding investigation is the calculation of the kinetic properties at low temperature [35], like relaxation or folding times. A major problem of MC simulations at low temperatures is that the ratio of rejected moves is very large. Here, the efficiency of the algorithm is increased at low temperature, using a Bortz-Kalos-Lebowitz (BKL) type algorithm [40,29], adapted to lattice simulations using MC*, noted BKL*. The probability to reject a move from the conformation $m$ during one step is noted $w_m$ and using Eqs. (4), (5), and (16), one obtains

$$w_m=1-\sum_{n\neq m}W(m\rightarrow n) \tag{21}$$

$$=1-\sum_{n\neq m}W_0(m\rightarrow n)a(m\rightarrow n) \tag{22}$$

$$=1-\frac{1}{2N-5}\sum_{n\neq m}\frac{\delta^{mn}}{1+\exp(\Delta E_n^m/T)}. \tag{23}$$

The probability $P_m(k)$ to accept a move from the conformation $m$ after exactly $k$ MC* steps is the product of the probability to reject a move during $k-1$ MC* steps by the probability to accept any move during one step,

$$P_m(k)=w_m^{k-1}(1-w_m). \tag{24}$$

Obviously, as $w_m<1$, the relations $\Sigma_{k=1}^{\infty}P_m(k)=1$, $\forall m$ are always satisfied. At each step of this algorithm a random integer number $k$ is chosen in the density of probability $P_m(k)$, then the conformation $m$ is counted $k$ times for the statistically averaged calculations and a move $m\rightarrow n$ chosen with the following normalized probability:

$$t(m\rightarrow n)=\frac{W(m\rightarrow n)}{\sum_{n'\neq m}W(m\rightarrow n')} \tag{25}$$

$$=\frac{\dfrac{\delta^{mn}}{1+\exp(\Delta E_n^m/T)}}{\sum_{n'\neq m}\dfrac{\delta^{mn'}}{1+\exp(\Delta E_{n'}^m/T)}} \tag{26}$$

is performed at the step $k+1$ and the conformation $n$ becomes the new current conformation.

Figure 7 shows the folding times obtained by using BKL* at low temperature with three different simulations, following the choice of the first conformations set. The folding times ($t_{\mathrm{fold}}$) are defined as the average over 500 trajectories of the number of MC* steps needed to reach the lowest energy conformation (shown in Fig. 8).

The simulation "$T$" for which the trap structure is chosen as the first conformation of the trajectories. The trap conformation is defined as the one that presents the highest energy barrier to reach the native state (see inset of Fig. 7). The trap has been calculated by solving the master equation of the system with the choice of $r$ made in this work. This structure (shown in Fig. 8) is the same as that found in a previous work [41].

The simulation "$E$" for which the first conformations are chosen at random among the extended ones, i.e., conformations without any contact.

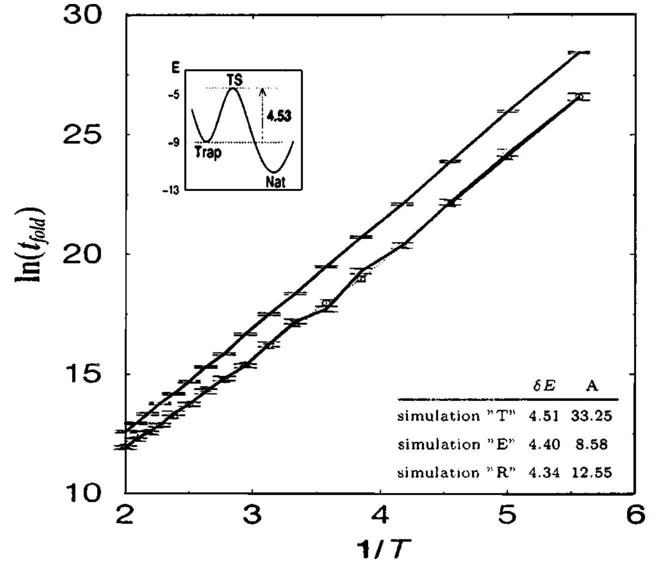

FIG. 7. Main plot: logarithm of the mean folding time versus the inverse of the temperature for the simulation "$T$" (circle), "$R$" (square), "$E$" (triangle). The error bars are one standard deviation about the mean. Plot in inset: schematic pathway from the trap to the native (Nat) conformation through the transition state (TS) function of the energy. Table in inset: value of the parameters $\delta E$ and $A$ of the Arrhenius laws $t_{\text{fold}}(T)=A\exp(\delta E/T)$ for the "$T$," "$E$," and "$R$" simulations (see text). The regressions are performed over the points on the solids lines (low temperature).

The simulation "$R$" that starts with conformations chosen at random among the whole conformational space.

The energy difference between the trap and the transition state equals $\Delta E=4.53$ (inset of Fig. 7). In each of the three simulations, Arrhenius law is recovered, $t_{\text{fold}}(T)=A\exp(\delta E/T)$ at low temperature ($T=0.24$, 0.22, 0.20, 0.18), for the folding times computed with the MC* method. The results of $\delta E$, shown in the table in the inset of Fig. 7, are in very good agreement (less than 1% for the $T$ simulation) with the value of $\Delta E$ and strongly support the proposed method for the calculation of Arrhenius law parameters.

## IV. DISCUSSION

In the following, we will focus on the properties of the conformational rigidity $w_m^{(0)}$ and on their role during the folding processes.

During the folding time, the chain is not at thermal equilibrium with the solvent bulk. At low temperature, only the native conformation is relevant; however, it does not appear during the folding time except at the ultimate step, as the end of the folding time is given by $n^{\text{Nat}}(t_{\text{fold}})=1$. This period can be viewed as the first stage of the process of convergence towards thermal equilibrium.

During the second phase, the system evolves towards thermal equilibrium by increasing the occurrence of the native state, until a good accordance with the equilibrium probability is found. Then, the relaxation time can be defined as the time $t_{\text{rel}}$ that satisfies $\pi^{\text{Nat}}(t_{\text{rel}})=n^{\text{Nat}}(t_{\text{rel}})/t_{\text{rel}}=P_{\text{eq}}^{\text{Nat}}$. During this relaxation period, the system evolves towards the thermal equilibrium, but conformations of the chain do not occur in proportion to their Boltzmann weights.

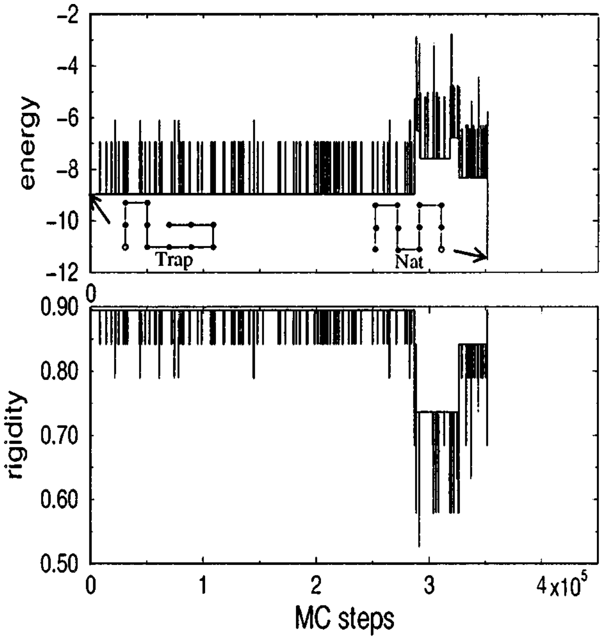

FIG. 8. Energy (top) and rigidity (bottom) versus the MC* steps of a typical trajectory of folding simulation at $T=0.4$, starting with the trap conformation. The trap and the native (Nat) conformations are shown. The empty circle is for the first monomer.

During the third and last phase that occurs after the relaxation time, all the structures occur with a probability proportional to their Boltzmann weights, $\pi^{(m)}=P_{\text{eq}}^{(m)}$.

However, the biological function and activity of a protein are closely related to the shape of its native conformation. Then, from a biological point of view, the folding time is a more meaningful quantity than the relaxation time, because it characterizes the time needed to reach the biologically active conformation.

To study the kinetic path followed by the chain, MC* is a powerful and well adapted method, even during the folding or the relaxation times of the chain. The probability to perform a move is a function of the difference of energy between the connected structures and of the temperature. On the other hand, the *a priori* probability to attempt a move is related to the rigidity of each conformation. Then, it does not depend on the temperature, whereas the probability to accept the move is temperature dependent. This fundamental difference between the two mechanisms involved in the procedure implicates that the kinetic path of folding presents qualitative differences depending on the temperature of the simulation.

In order to understand the mechanism of folding in the range of temperature where the protein is biologically active, i.e., at medium temperature, the kinetic properties of the chain at high and low temperatures are first described independently below. These extremum cases are of theoretical interest, but also give an insight into of the property of the chain at medium temperature.

At high temperature, following Eq. (5), the acceptance ratio is $a(m\rightarrow n)=0.5$ for all the connections $m\rightarrow n$. Then, the evolution of the chain is only guided by the values of the rigidity of the conformations. When a structure that presents a large number of connections, i.e., having small values of $w_0^{(m)}$ occurs, a new transition takes place very quickly. On the contrary, conformations with very few connections are rarely reached, but when such structures occur, many MC* steps are needed before trying a move. However, in any case, for long simulations, all the conformations have the same probability to occur and they are counted during the same number of MC* steps. The less rigid structures are more often reached, but a move is accepted quicker than from more rigid structures. Then, at high temperature, the mechanism that guides the chain towards the native structure is simply a random walk process that takes into account the rigidity of the conformations. This random process is particular because whereas all the structures have the same probability to occur on an average, they have not the same probability to be reached during the MC* simulation. Moreover, the average time spent, at each time that a given conformation is reached, is also dependent on the conformation. The chain finds the native conformation following this random walk adapted to lattice model. Omission of rigidity parameters would lead to an increase in the probability of occurrence of the extended conformations and would give a wrong view of the mechanism of the kinetics of folding at high temperature.

However, the case of the folding process at high temperature only presents a theoretical interest because, in this case, the chain is always in a denatured phase and when the native state is reached, it is only for a very short time and then the molecule is not really biologically active.

At low temperature, the acceptance ratio is very selective. When temperature goes to zero, $a(m\rightarrow n)\rightarrow 0$ if $\Delta E>0$ and $a(m\rightarrow n)\rightarrow 1$ if $\Delta E<0$, and the acceptance ratio plays a more important role in the selection of the transitions than in the previous case. The rigidities of the conformations still play a role, but minor, in the folding process at low temperature. However, an accepted transition cannot lead to an increase in the energy and all the moves undertaken towards conformations of lower energy have the same probability to be performed because $a_{\Delta E<0}(m\rightarrow n)\rightarrow 1$. The chain is mostly trapped in local minima and the kinetics of folding is very slow.

In nature, at very low temperature, the solvent around protein is converted into ice. Then, kinetics of protein folding is simply frozen by the crystallization of the solvent. The potential used in this work is temperature independent and does not well mimic the effect of the solvent at low temperatures. Then, investigations on the mechanisms of the protein folding at very low temperature remain illusive with such potential. However, this study, as well as high-temperature simulations, present theoretical approaches very useful to understand the folding of the proteins at medium temperature.

At medium temperature, the evolution of the chain towards its native structure is dictated by both rigidities and energy differences. In order to understand the folding processes, many kinetic paths beginning with the trap conformation and ending with the native structure have been computed at $T=0.4$, using MC* simulations. They all exhibit similar properties and a typical trajectory is shown in Fig. 8. The native and the trap conformations have very low energy ($E_{\mathrm{Nat}}=-11.5031$, $E_{\mathrm{Trap}}=-8.9627$), and are also very compact structures (five contacts for the trap and six for the native conformation), and then, they are very rigid ($w_{\mathrm{Nat}}^{(0)}=w_{\mathrm{Trap}}^{(0)}=0.894$). As the trap is the first conformation of each trajectory, no transition is accepted during a lot of MC* steps. As the allowed transitions are only a local modification of the chain, when a move is selected and accepted, the chain occurs in a new conformation of low energy and of still relatively high compacity and then high rigidity. As the transition of the highest probability from this conformation is the way back to the trap conformation, there are oscillations of very low frequencies in the lower part of the conformational valley of the trap between the trap and the few conformations connected to it. The frequency of the accepted moves, computed over the last $N_f$ moves (here $N_f=20$), is defined as $N_f$ divided by the number of MC* steps needed to accept the last $N_f$ moves and is shown in Fig. 9. These moves of low frequencies in the trap valley occur during the $2.7\times10^5$ first steps, afterwards a conformation among the huge set of transition state is reached and permits to escape from this conformational valley. The transition states, which exhibit common properties (few intrachain contacts, great flexibilities, and high energies), have a small equilibrium probability of occurrence, whereas they are easily accessible at a topological point of view as they are not very rigid. On the other hand, the mean time of occurrence of these conformations is very short as they are very flexible. Then, the chain evolves from valley to valley following this mechanism. As the ground states of the other valleys have smaller rigidities and higher energies than the trap of the system, the oscillations are of higher frequencies. When the main valley, i.e., the valley of the native state, is reached, the frequencies of transitions are very high. In this ultimate funnel, the chain folds towards the native structure by minimizing its energy. But, as the chain is driving towards the native structure, its conformations becomes more and more compact. The folding pathway admits then less and less possibilities of connections

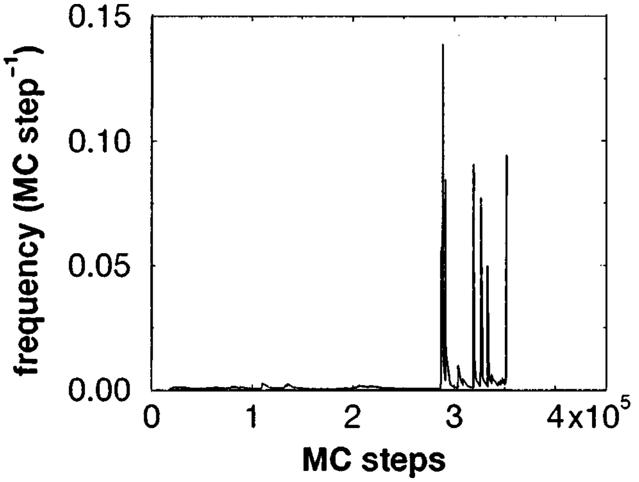

FIG. 9. Frequency of the accepted moves, computed on the 20 last accepted transitions (see text), in the MC* simulation at $T=0.4$ as a function of the MC* steps.

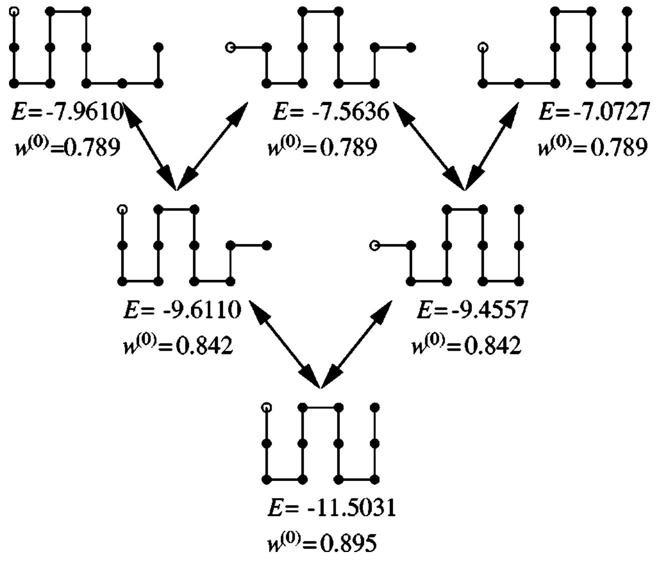

FIG. 10. Last connections between the conformations leading to the native structure.

between conformations that are more and more rigid, as can be seen on the last part of the connection graph shown in Fig. 10. However, most of the attempted moves are accepted and they guide the chain towards the lower part of the funnel in a relatively few steps. In other words, the kinetics of folding is slowed down as the chain goes down the funnel, although it evolves harshly towards the native structure, with a relatively high accepted move frequency compared to the accepted move frequencies in others valleys.

Ten simulations have been carried out at the same temperature ($T=0.4$) and the logarithm of the occurrence of conformations with energy $E$ during the folding time period versus the energy has been computed and plotted in Fig. 11. In a simulation performed at thermal equilibrium with a thermostat, a linear relation between both quantities would have been found. It appears clearly that during the folding time, only the part of the spectrum above the dashed line is well sampled. A linear regression of this region of the spectrum gives $n(E)=A\exp(-E/T^*)$ with $T^*=0.48$. The value of $T^*$ is close to the temperature of the simulation, i.e., $T=0.40$, showing that the subspace above the dashed lines is at thermal equilibrium with the bath. The coefficient $A$ is not the inverse of the partition function of the whole system because the conformations of low energies are badly sampled. It can be seen as the inverse of a "reduced partition function" computed on the subspace of conformations at equilibrium with the bath temperature. The conformations of this region belong to valleys other than that of the native state and they occur with probabilities larger than, but proportional to, their equilibrium probabilities. Then, at this stage it can be assumed that the subsystem consisting of all the valleys, except that of the native state, is at thermal equilibrium. The time needed to reach this partial equilibrium corresponds to

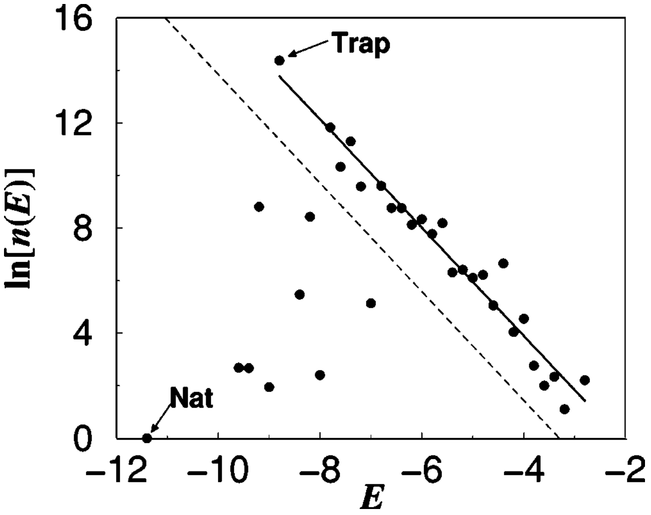

FIG. 11. Logarithm of the occurrence of the energy $E$ during the folding times of ten trajectories.

the folding time defined above. The conformations of low energy which are undersampled will mostly occur during the second stage of the relaxation, appearing for times larger than $t_{\mathrm{fold}}$ and smaller than $t_{\mathrm{rel}}$ by performing a sampling mainly localized in the native valley. For $t \gg t_{\mathrm{rel}}$, each conformation occurs with its equilibrium probability, as is shown in Fig. 6.

## V. CONCLUSION

The results presented in this work emphasize that the proposed MC* method is well adapted to study the dynamics of protein folding. It has been shown that not only the difference of energies between the conformations but also the rigidity of the conformations have to be taken into account in the MC* simulation in order to sample correctly the conformational space. Moreover, the BKL algorithm has been implemented and would be a good technique to provide low-temperature studies and rollover behavior [42] observed for small single domain proteins for which the folding arm of the chevron plot is not linear under native conditions.

Kinetic paths have been studied and some general features to give an insight into the mechanism that drives a protein towards its native structure at medium temperature. During the folding time of this process, only a part of the conformational space is sampled in proportion to the Boltzmann weights of the conformations. Afterwards, between the folding and relaxation times, the occurrence of the conformations of low energy increases and all the conformations occur in proportion to their Boltzmann weights and the thermal equilibrium of the chain is reached.

Finally, this method had been applied only to a short chain in order to check its efficiency, but it is easily applicable to a longer chain on a two- or three-dimensional lattice.

[1] C.B. Anfinsen, E. Haber, M. Sela, and F.H. White, Proc. Natl. Acad. Sci. U.S.A. **47**, 1309 (1961).

[2] C. Levinthal, J. Chim. Phys. Phys.-Chim. Biol. **65**, 44 (1968).

[3] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).

[4] C.L. Brooks III, M. Karplus, and B.M. Pettitt, *Proteins* (Wiley, New York, 1988).

[5] K.D. Gibson and H.A. Scheraga, Physiol. Chem. Phys. **1**, 109 (1969).

[6] N. Gõ and H.A. Scheraga, J. Chem. Phys. **51**, 4751 (1969).

[7] B. Lee and F.M. Richards, J. Mol. Biol. **55**, 379 (1971).

[8] C. Chothia, Nature (London) **248**, 338 (1974).

[9] H. Zhang, C.F. Wong, T. Thacher, and H. Rabitz, Proteins **23**, 218 (1995).

[10] O. Collet and S. Premilat, J. Mol. Struct.: THEOCHEM **363**, 151 (1996).

[11] O. Collet, S. Premilat, B. Maigret, and H.A. Scheraga, Biopolymers **42**, 363 (1997).

[12] K.F. Lau and K.A. Dill, Macromolecules **22**, 3986 (1989).

[13] E.I. Shakhnovich and A.M. Gutin, Biophys. Chem. **34**, 187 (1989).

[14] H.S. Chan and K.A. Dill, J. Chem. Phys. **92**, 3118 (1990).

[15] E.I. Shakhnovich and A.M. Gutin, Nature (London) **346**, 773 (1990).

[16] H.S. Chan and K.A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).

[17] A. Šali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[18] A. Šali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).

[19] H.S. Chan and K.A. Dill, J. Chem. Phys. **100**, 9238 (1994).

[20] H. Kaya and H.S. Chan, Proteins: Struct., Funct., Genet. **40**, 637 (2000).

[21] A. Dinner, A. Šali, M. Karplus, and E. Shakhnovich, J. Chem. Phys. **101**, 1444 (1994).

[22] H.S. Chan and K.A. Dill, J. Chem. Phys. **99**, 2116 (1993).

[23] E.I. Shakhnovich and A.M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).

[24] E.I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[25] E.I. Shakhnovich and A.M. Gutin, J. Chem. Phys. **93**, 5967 (1990).

[26] H.S. Chan and K.A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).

[27] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[28] K. Binder, *Monte Carlo Methods in Statistical Physics* (Springer, Berlin, 1979).

[29] M.E.J. Newman and G.T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, Oxford, 1999).

[30] R. Miller, C.A. Danko, M.J. Fasolka, A.C. Balazs, H.S. Chan, and K.A. Dill, J. Chem. Phys. **96**, 768 (1992).

[31] M. Karplus and E. Shakhnovich, *Protein Folding*, edited by T.E. Creighton (Freeman, New York, 1992), pp. 127–193.

[32] P.E. Leopold, M. Montal, and J.N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).

[33] N.D. Socci and J.N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).

[34] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).

[35] A. Gutin, A. Šali, V. Abkevich, M. Karplus, and E. Shakhnovich, J. Chem. Phys. **108**, 6466 (1998).

[36] N.D. Socci, J. N. Onuchic, and P.G. Wolynes, Proteins: Struct., Funct., Genet. **32**, 1136 (1998).

[37] C.M. Dobson, A. Sali, and M. Karplus, Angew. Chem., Int. Ed. Engl. **37**, 868 (1998).

[38] J.M. Sorenson and T. Head-Gordon, Fold Des **3**, 523 (1998).

[39] T.X. Hoang and M. Cieplak, J. Chem. Phys. **109**, 9192 (1998).

[40] A.B. Bortz, M.H. Kalos, and J.L. Lebowitz, J. Comput. Phys. **17**, 10 (1975).

[41] M. Cieplak, M. Henkel, J. Karbowski, and J.R. Banavar, Phys. Rev. Lett. **80**, 3654 (1998).

[42] H. Kaya and H.S. Chan, J. Mol. Biol. **315**, 899 (2000).