PHYSICAL REVIEW E **76**, 026113 (2007)

Uncovering the topology of configuration space networks

David Gfeller,[1] David Morton de Lachapelle,[1] Paolo De Los Rios,[1] Guido Caldarelli,[2,3] and Francesco Rao[2]
[1]*Laboratoire de Biophysique Statistique, SB/ITP, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland*
[2]*Museo Storico della Fisica e Centro Studi e Ricerche "E. Fermi," 00184 Rome, Italy*
[3]*SMC, INFM-CNR, Dipartimento di Fisica, Università "Sapienza," Piazzale Aldo Moro 5, 00185 Rome, Italy*
(Received 18 April 2007; published 27 August 2007)

The configuration space network (CSN) of a dynamical system is an effective approach to represent the ensemble of configurations sampled during a simulation and their dynamic connectivity. To elucidate the connection between the CSN topology and the underlying free-energy landscape governing the system dynamics and thermodynamics, an analytical solution is provided to explain the heavy tail of the degree distribution, neighbor connectivity, and clustering coefficient. This derivation allows us to understand the universal CSN topology observed in systems ranging from a simple quadratic well to the native state of the beta3s peptide and a two-dimensional lattice heteropolymer. Moreover, CSNs are shown to fall in the general class of complex networks described by the fitness model.

DOI: 10.1103/PhysRevE.76.026113 PACS number(s): 89.75.Fb, 89.75.Da, 87.15.He

I. INTRODUCTION

The use of complex networks and graph theory to describe complex systems ranging from the worldwide web (WWW) to protein interaction networks is by now well established (different books and reviews are available about this topic [1–5]). A large class of these systems attain complexity by means of their internal dynamics [6], which is often revealed by computer simulations. One example of such systems is the folding of proteins, for which simulations have been extensively used in structural biology. Nowadays several molecular dynamics (MD) packages (CHARMM [7], GROMACS [8], AMBER [9]) are available to probe in real time the dynamics of folding and unfolding. However, because of the large number of degrees of freedom [10] involved in the process, the results of MD simulations form in themselves a highly complex system. As a consequence, a detailed, unbiased description of the free-energy landscape underlying the thermodynamics and kinetics cannot easily be extracted.

To tackle this complexity, new approaches based on complex networks have recently been introduced, showing that a network description is effective for the analysis and visualization of simulation results. In [11], for instance, the topology of the configurations of a short lattice polymer has been mapped onto a network. Doye and Massen have also applied graph analysis to study the organization of the potential energy minima in a Lennard-Jones cluster of atoms [12,13]. In another work, the concept of disconnectivity graphs has been used to analyze the free energy of a tetrapeptide and a β-hairpin [14,15]. Finally, the free-energy landscape of a three-stranded β-sheet (beta3s) and alanine dipeptide sampled by MD simulations have been represented as a configuration space network (CSN) [16,17].

Given the time evolution of a dynamical system, the CSN represents the ensemble of microstates (configurations) sampled during a simulation, and their dynamic connectivity. In this representation, nodes are system configurations and links are direct transitions between the configurations sampled during the simulation. The CSN topology shares several features with other networks representing systems as different as cell's functional architecture [18], scientific collaborations [19], and the WWW [20]. In particular, it has been shown [16] that the degree distribution of the beta3s peptide CSN exhibits a heavy tail well approximated by a power-law, and a disassortative behavior for the average neighbor connectivity distribution. Moreover, the clustering coefficient presents a decay compatible with a $1/k$ function for large values of the degree, which has been interpreted as the presence of a hierarchical organization of the nodes [21]. Recently, the connection between the CSN topological clusters and free-energy basins has been explored, and an analytical solution for the node weight distribution observed in CSNs has been provided with the help of simple energy landscape models [17]. Following these lines of research, the challenge is now to find the connection between network topology, system dynamics, and free-energy landscape organization. In this work, we focus on the degree distribution $P(k)$, the average neighbor connectivity $K_{nn}(k)$, and the clustering coefficient $C(k)$ observed in CSNs. Several studies [22–25] have shown that the analysis of the above three distributions is an important step toward the understanding of the network organization and architecture. The results presented below provide a rationale for the origin of several unexplained properties of CSNs.

The paper is organized as follows. Section II describes in detail how CSNs are built. Section III shows how the degree distribution, the average neighbor degree, and the clustering coefficient relate to the free-energy landscape. In Sec. IV, an analytical derivation and simulation results are presented for the quadratic well model. Then the CSNs obtained from beta3s peptide and lattice heteropolymer simulations are analyzed in Sec. V. Finally, the connection between CSNs and the fitness model is discussed in Sec. VI and conclusions are presented in Sec. VII.

II. CONFIGURATION SPACE NETWORKS

The simulation of a dynamical system, like a peptide or a protein, results in a time series of snapshots representing the dynamics. The CSN of this kind of process gives a synthetic

view of the configurations and transitions observed during the simulation. System configurations are the nodes, and a link is placed between two nodes if they appear consecutively in the time series. The time step between two snapshots, t_s (usually called the configuration saving time), is a free parameter: $t_s = i_t M$, where i_t is the integration time step for the simulation and M is the number of microscopic steps between two snapshots. When M approaches 1, only configurations spatially close to each other are connected together. Therefore a link is a temporal relation between configurations, and changing M changes the set of links.

The weight of a link w_{ij} represents the number of direct transitions from node i to node j. Similarly, the weight w_i of a node is given by the number of times configuration i has been visited. The weight distribution of CSN has been discussed in previous work [17].

The degree of a node is defined as the number of links including self-loops, corresponding to the number of configurations accessed in M steps during the dynamics. Because of finite-time simulations, the CSN is a directed network: If the system visited node j M steps after node i, the converse is not automatically true. Hence k^{in} and k^{out} are not always equal. However, the asymmetry of the links is weak for two reasons. First, the simulation is run long enough to almost ensure $w_{i\to j} = w_{j\to i}$ (which is in fact equivalent to detailed balance). Second, the total weight of the incoming links has to be equal to the total weight of the outgoing links by construction of the network.

In the following, the degree of a node, k_i, is defined as the out-degree k_i^{out}. Similarly, the average neighbor degree $k_{NN}(k)$ is the average out-degree of the neighbors of the nodes with degree k. The out-degree correlation between connected nodes is further characterized by the assortativity coefficient q [26]. Finally, the clustering coefficient is computed as the total number of three-steps cycles (triangles) starting at node i (N_i^{Δ}), divided by the maximum number of three-step cycles one can have in the considered graph: $c_i = N_i^{\Delta}/k_i^{out}k_i^{in}$.

III. ANALYTICAL APPROACH

As already mentioned, the main objective is to understand the relation between the network topology and the free-energy landscape. Unfortunately, even the degree of a node cannot be easily computed from the knowledge of the energy landscape for any M. For this reason, we restricted ourselves to large values of M corresponding to a random sampling of the landscape (uncorrelated exploration). In this case an analytical approach can be carried out. Let us consider the free-energy landscape $U(\mathbf{x})$ (in k_BT units). The probability density on the free-energy landscape is given by $W(\mathbf{x}) = W_0 \exp[-U(\mathbf{x})]$. The system configurations (i.e., CSN nodes) are defined as hypercubic cells of size a^D, where D is the dimension of the $\mathbf{x}$ space. Assuming that a is chosen small enough such that $\exp[-U(\mathbf{x})]$ is almost constant on each cell, the probability to visit a configuration at $\mathbf{x}_1$ at a given time is $P(\mathbf{x}_1) = a^D W_0 \exp[-U(\mathbf{x}_1)]$, and the expected number of times two configurations at position $\mathbf{x}_1$ and $\mathbf{x}_2$ are visited consecutively is given by

$$W(\mathbf{x}_1,\mathbf{x}_2) = W_N e^{-U(\mathbf{x}_2)-U(\mathbf{x}_1)} \tag{1}$$

where $W_N = Na^{2D}W_0^2$, and N is the total number of snapshots. The expression above predicts link weights. To compute the degree, the quantity of interest is the probability $P(\mathbf{x}_1,\mathbf{x}_2)$ to have a link between two configurations (no matter how often the link has been visited). Assuming that the probability distribution of visiting s times the node at $\mathbf{x}_1$ is peaked around its average value $P(\mathbf{x}_1)$, $P(\mathbf{x}_1,\mathbf{x}_2)$ is evaluated as one minus the probability to have no links:

$$P(\mathbf{x}_1,\mathbf{x}_2) = 1 - [1 - P(\mathbf{x}_2)]^{NP(\mathbf{x}_1)} \approx 1 - e^{-NP(\mathbf{x}_1)P(\mathbf{x}_2)}, \tag{2}$$

where the second equality holds in the limit of small $P(\mathbf{x}_2)$, which is true if the number of configurations is large. $NP(\mathbf{x}_1)$ is the expected number of times the configuration at $\mathbf{x}_1$ has been visited, and $1-P(\mathbf{x}_2)$ is the probability not to visit the configuration at $\mathbf{x}_2$. Equation (2) is indeed an approximation. An exact expression would require to sum up the probability of visiting s times node $\mathbf{x}_1$ multiplied by the probability of never visiting $\mathbf{x}_2$ right after $\mathbf{x}_1$, i.e., $[1-P(\mathbf{x}_2)]^s$, excluding the cases in which $\mathbf{x}_1$ has been visited several times consecutively. However, it is very difficult to express this in a simple form, and approximations are required to proceed further with analytical calculations. From Eq. (2), two asymptotic behaviors can be derived.

(1) If $NP(\mathbf{x}_2)P(\mathbf{x}_1)$ is large, then $P(\mathbf{x}_1,\mathbf{x}_2) \approx 1$. This is the *saturation regime* since $\mathbf{x}_1$ and $\mathbf{x}_2$ are almost certainly connected.

(2) If $NP(\mathbf{x}_2)P(\mathbf{x}_1)$ is small, the *sparse regime* is reached, which describes low-probability connections. In this regime an nth-order expansion is meaningful:

$$P^{(n)}(\mathbf{x}_1,\mathbf{x}_2) = \sum_{j=1}^{n} \frac{1}{j!}[NP(\mathbf{x}_1)P(\mathbf{x}_2)]^j(-1)^{j+1}. \tag{3}$$

The first term in Eq. (3) is equal to $W(\mathbf{x}_1,\mathbf{x}_2)$ in Eq. (1). Taking only the first term in the sum corresponds to the case in which links are distributed avoiding as much as possible the presence of double links. If $NP(\mathbf{x}_2)P(\mathbf{x}_1) \ll 1$, it provides a good approximation of the sum, while for $NP(\mathbf{x}_2)P(\mathbf{x}_1) \approx 1$, it slightly overestimates the real probability to have a link between two nodes. Applying the considerations above, an approximation of $P(\mathbf{x}_1,\mathbf{x}_2)$ is given by

$$E^{(n)}(\mathbf{x}_1,\mathbf{x}_2) = \min[1, P^{(n)}(\mathbf{x}_1,\mathbf{x}_2)]. \tag{4}$$

Equation (4) defines the probability to have a link between two nodes, depending only on a parameter associated with each node [in this case the energy $-U(\mathbf{x})$]. Such systems have been previously described in the fitness model framework [27–29] (see Sec. VI). The degree of a node at $\mathbf{x}$, its average neighbor degree, and the expected number of triangles the node is part of are then given by the following three expressions:

$$k(\mathbf{x}) = \frac{1}{a^D}\int_V d\mathbf{x}_1 E^{(n)}(\mathbf{x},\mathbf{x}_1), \tag{5}$$

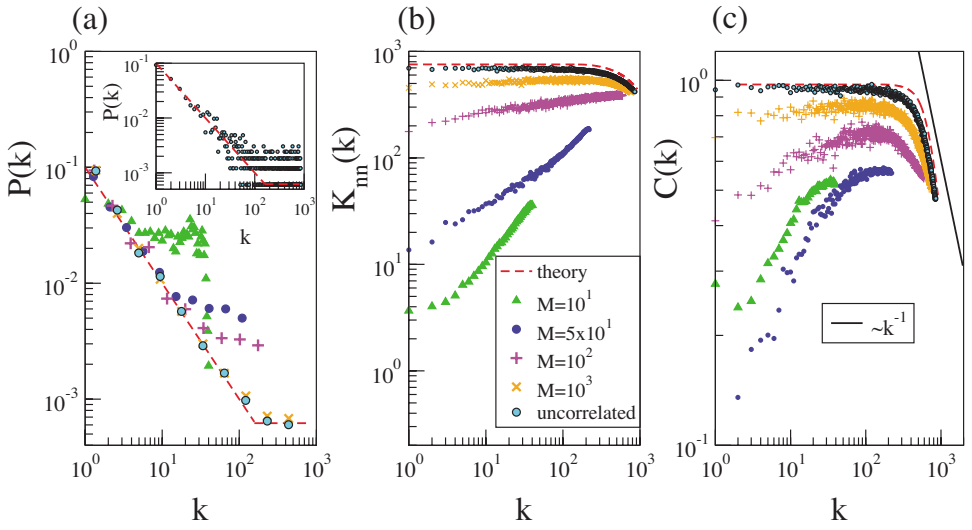

FIG. 1. (Color online) Network topology for the quadratic well in $D=2$ dimensions and different values of the parameter M. (a) Degree distribution. For clarity a binning has been applied for $M>10$. Inset: Degree distribution for uncorrelated sampling without binning. (b) Average neighbor degree. (c) Clustering coefficient. Blue circles surrounded by black correspond to a random sampling of the energy landscape (uncorrelated case; see text). Red dashed line shows the analytical estimation.

$$K_{nn}(\mathbf{x}) = \frac{1}{a^D}\frac{1}{k(\mathbf{x})}\int_V d\mathbf{x}_1 E^{(n)}(\mathbf{x},\mathbf{x}_1)k(\mathbf{x}_1), \tag{6}$$

$$N^{\triangle}(\mathbf{x}) = \frac{1}{a^{2D}}\int_V\int_V d\mathbf{x}_1 d\mathbf{x}_2 E^{(n)}(\mathbf{x},\mathbf{x}_1)E^{(n)}(\mathbf{x},\mathbf{x}_2)E^{(n)}(\mathbf{x}_1,\mathbf{x}_2). \tag{7}$$

Finally, assuming that the continuous approximation is valid and that the degree distribution of a node at $\mathbf{x}$ is peaked around its average value $k(\mathbf{x})$, the degree distribution reads

$$P(k) \sim \int_V d^D\mathbf{x}\ \delta(k-k(\mathbf{x})). \tag{8}$$

Inverting Eq. (5) and inserting it into Eqs. (6) and (7) gives the average neighbor connectivity $K_{nn}(k)$ and the clustering coefficient $C(k)$, respectively,

$$K_{nn}(k) = \frac{1}{a^D}\frac{1}{k}\int_V d\mathbf{x}_1 E^{(n)}(\mathbf{x}(\mathbf{k}),\mathbf{x}_1)k(\mathbf{x}_1), \tag{9}$$

$$C(k) = \frac{1}{k^2 a^{2D}}\int_V\int_V d\mathbf{x}_1 d\mathbf{x}_2 \times E^{(n)}(\mathbf{x}(\mathbf{k}),\mathbf{x}_1)E^{(n)}(\mathbf{x}(\mathbf{k}),\mathbf{x}_2)E^{(n)}(\mathbf{x}_1,\mathbf{x}_2). \tag{10}$$

IV. QUADRATIC WELL

In general, the free-energy landscapes of real systems are extremely complex, so that even writing down a mathematical expression is often impossible. However, close to the minimum of a basin (corresponding to configurations visited several times), such systems can often be approximately described by means of a Taylor expansion of the potential, whose first term is harmonic.

Therefore the quadratic well is a good benchmark to understand more complex CSNs [17], in particular for nodes near the minimum of an energy basin. In two dimensions, the potential is given by

$$U(x,y) = \frac{1}{2}(x^2+y^2) = \frac{1}{2}r^2. \tag{11}$$

Using radial coordinates and introducing Eq. (11) in Eq. (4) with $n=1$ gives ($W_0 = 1/2\pi$)

$$P^{(1)}(\mathbf{x}_1,\mathbf{x}_2) = 1 \Leftrightarrow \frac{(2\pi)^2}{a^4 N} = \exp\left(-\frac{1}{2}(r_1^2+r_2^2)\right)$$
$$\Leftrightarrow r_1^2 = 2\ln\left(\frac{a^4 N}{(2\pi)^2}\right) - r_2^2 = B - r_2^2 \tag{12}$$

with $B = 2\ln[a^4 N/(2\pi)^2]$. Hence a necessary condition for $P^{(1)}(\mathbf{x}_1,\mathbf{x}_2)>1$ is that both $r_1<\sqrt{B}$ and $r_2<\sqrt{B}$. The degree distribution is then obtained from Eq. (8) (see the Appendix for the complete derivation):

$$P(k) \sim \begin{cases} \text{const} & \text{if } r<\sqrt{B} \Leftrightarrow k>2\pi/a^2, \\ 1/k & \text{if } r>\sqrt{B} \Leftrightarrow k<2\pi/a^2. \end{cases} \tag{13}$$

Note that the flat tail for $k>2\pi/a^2$ is an artifact of the continuous approximation in $D=2$. Analytical calculations, for instance in $D=4$ where they are still manageable (but somewhat tedious), show a decreasing behavior even for $k>2\pi/a^2$.

In the same way, the average neighbor connectivity is computed from Eq. (9) and the clustering coefficient from Eq. (10). Results are shown graphically on Fig. 1 (detailed calculations are included in the Appendix).

In order to compare the analytical predictions obtained above for $n=1$ with the CSN topology obtained from simulations, a Langevin dynamics with potential energy defined by Eq. (11) is performed according to the equation of motion

$$\gamma\dot{\mathbf{x}} = -\frac{\partial U}{\partial \mathbf{x}} + f(t),$$

where γ is the friction coefficient and $f(t)$ is a white noise with mean value $\langle f(t)\rangle = 0$ and $\langle f(t)f(t')\rangle = \delta(t-t')$. Without loss of generality, γ is set to 1 (this merely corresponds to a rescaling of the time) and the integration step to $i_t = 0.001$ used in the simulations. In the case of the two-dimensional quadratic well, configurations are defined as square cells of size a^2 ($a=0.2$). A total number of $N = 3\times 10^6$ time steps has been used in the simulations.

The degree distribution $P(k)$ for different values of the parameter M is shown in Fig. 1(a). The distribution follows a power-law of the form $1/k$ for values of the parameter $M \geqslant 100$. In this example, every CSN realization with $M \geqslant 10^4$ are equivalent to a random sampling of the energy landscape with probabilities given by $W_0 e^{-U(x)}$ (black points in the figure). Hence, for these values of M, the distribution follows the analytical prediction.

In Fig. 1(b), the average neighbor connectivity $K_{nn}(k)$ is plotted for different values of M. There is a change in the behavior of $K_{nn}(k)$ as M increases. For low values of M, $K_{nn}(k)$ is an increasing function of k, which indicates an assortative behavior. This is no longer true for large values of M. In this regime, $K_{nn}(k)$ shows a decaying tail characteristic of disassortative regime. For $M \geqslant 10^4$, the curves cannot be distinguished from the one obtained by uncorrelated sampling. The flat region for small k arises because nodes with low degree tend to connect to nodes with high degree which lie at the bottom of the basin. Indeed, for large M, transitions starting at a node far from the minimum are likely to end up at the bottom of the basin, which is characterized by nodes with a large degree. On the other hand, for small values of M, only neighbor configurations are visited consecutively.

As already pointed out, the approximation $n=1$ has the effect of slightly overestimating the node degree, which explains why results for the uncorrelated sampling are found to be below the analytical approximation. The assortativity coefficient q [26] for different values of M shows the same transition between assortative and disassortative regimes (see Fig. 2). For $M < 100$, the network presents a strong assortativity characterized by values around $q \approx 0.8$. Increasing the value of M makes the assortativity coefficient drop to values smaller than -0.3, indicating that the system has undergone an assortative-to-disassortative transition. Therefore, CSNs built from the same physical process, i.e., diffusion in a well, exhibit a changing assortativity [30]

In the same way, the clustering coefficient $C(k)$ exhibits different behaviors as a function of M. For $M < 1000$, the value of $C(k)$ grows, indicating that triangles easily form at the bottom of the basin. On the other hand, as M increases, $C(k)$ shows a decaying tail for large values of k. For $M > 10^4$, $C(k)$ obtained in the quadratic well follows the analytical prediction of Eq. (A4).

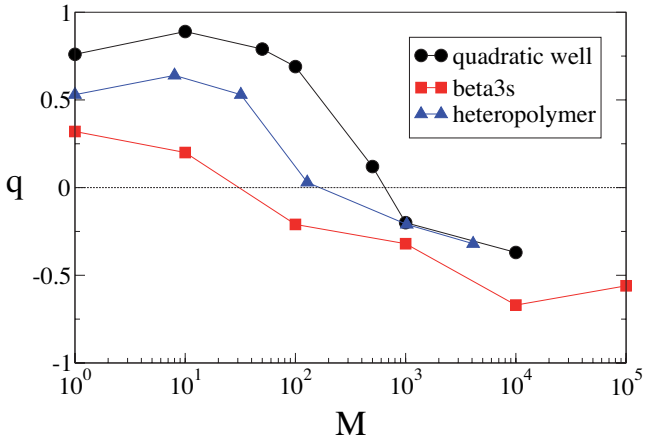

FIG. 2. (Color online) Assortativity coefficient for different values of the parameter M in the three systems under study.

The changing behavior of the CSN topology of a quadratic well for different values of the configuration saving time can be understood in a more general kinetic framework when considering the relaxation times to a given configuration of the landscape. In Fig. 3(a) the distribution of the relaxation times to the configuration lying at the bottom of the well for three different configurations is shown. The relaxation times starting from configurations close to the bot-

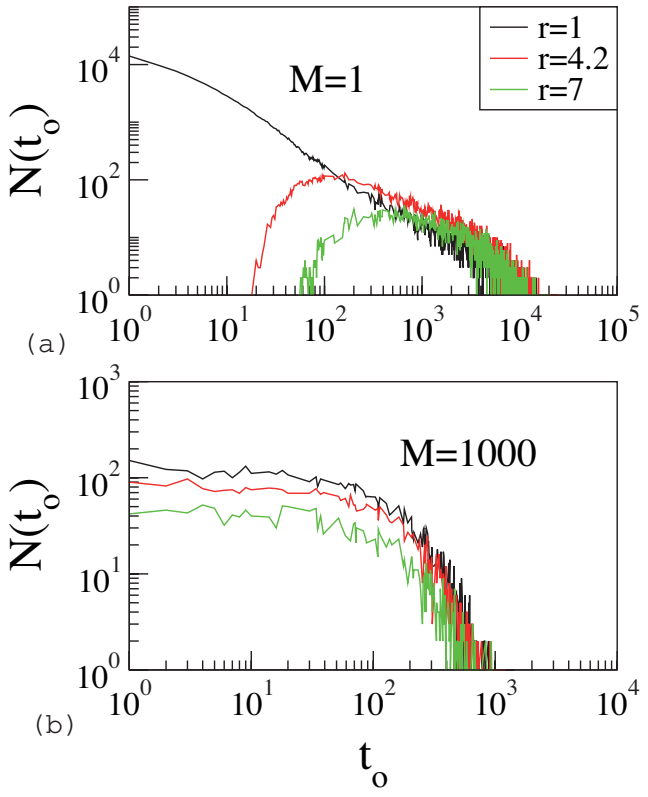

FIG. 3. (Color online) Distribution of the relaxation times from three different initial configurations to the bottom configuration of the quadratic well in $D=2$ dimensions for $M=$(a) 1 and (b) 1000. The value r indicates the radial distance from the starting node for each of the three curves.

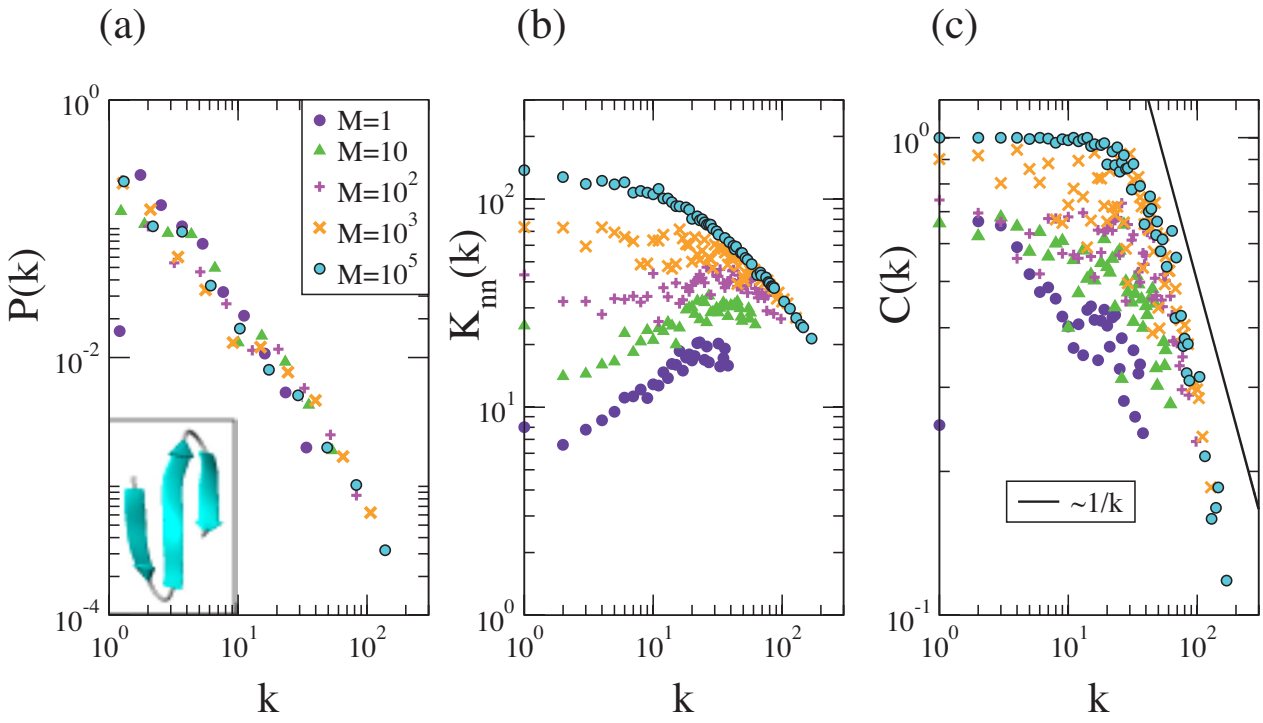

FIG. 4. (Color online) Network topology for the beta3s peptide CSN at different values of the parameter M. (a) Degree distribution. To reduce noise a logarithmic binning has been applied. The native state of beta3s is displayed in the inset. (b) Average neighbor connectivity. (c) Clustering coefficient.

tom node (small r) exhibit a downhill distribution, which is not the case for larger values of r. However, as M increases, all distributions overlap (up to a global multiplicative factor) indicating that the kinetics to the bottom is the same for all configurations [see Fig. 3(b)]. This corresponds to the uncorrelated regime of Eq. (1). In this case, the probability to have a link between two configurations depends only on the configuration weight.

V. NATIVE STATE OF A TRIPLE STRANDED β–SHEET AND A LATTICE HETEROPOLYMER

The analytical and numerical results obtained above are crucial for a correct interpretation of the CSN topology observed in complex systems for which direct application of Eq. (8)–(10) is unfeasible. In the following, the CSN topology of the native basin of a triple stranded β-sheet peptide (beta3s) sampled by MD as well as of a lattice heteropolymer sampled by Monte Carlo simulations are investigated (see Figs. 4 and 5).

The MD simulation of the native state of beta3s is performed at 270 K for a total of 10 ns, which is enough for the correct sampling of the basin. The low temperature prevented the system from jumping to a different basin. The MD simulation is performed using the CHARMM PARAM19 force field [7] and an integration time step of i_t=2 fs. A mean field approximation based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute [31]. The two parameters of the solvation model were optimized without using beta3s. The same force field and implicit solvent model have been used recently in MD simulations of various systems [32–34].

The secondary structure is worked out [35] for each snapshot saved along the MD trajectory. Here a configuration (i.e., a CSN node) is defined as a single string of secondary structure e.g., the most populated configuration for beta3s at 270 K (see inset of Fig. 4) is -EEE-STTEEEEESSEEEE-. The total number of 5×10^6 snapshots sampled during the MD simulation resulted in 249 secondary structure configurations. There are eight possible letters in the secondary structure "alphabet": H, G, I, E, B, T, S, and -, standing for α-helix, 3_{10}–helix, π-helix, extended, isolated β-bridge, hydrogen-bonded turn, bend, and unstructured, respectively. Since the N- and C-terminal residues are always assigned an -, a 20 residue peptide can, in principle, assume $8^{18}\approx10^{16}$ configurations.

The two-dimensional lattice heteropolymer is simulated in the framework of the popular hydrophobic-polar (HP) model [36–40]. In this description, the amino acid sequence of a protein is represented as a binary sequence of hydrophobic and polar residues. The results presented here are obtained with the random sequence HHPHPPHHPPHHPH (inset of Fig. 5). Note that similar results are observed for different HP sequences (random and proteinlike), as well as for different numbers of residues, ranging from 10 to 20 (a detailed presentation of these results is in preparation). The time series of configurations is generated from moves of the polymer according to the Metropolis rule. It is important to mention that the qualitative observations do not depend on the set of moves (local moves and the global "pivot" moves [41,42] have been tested). In the standard HP model, the energy of a

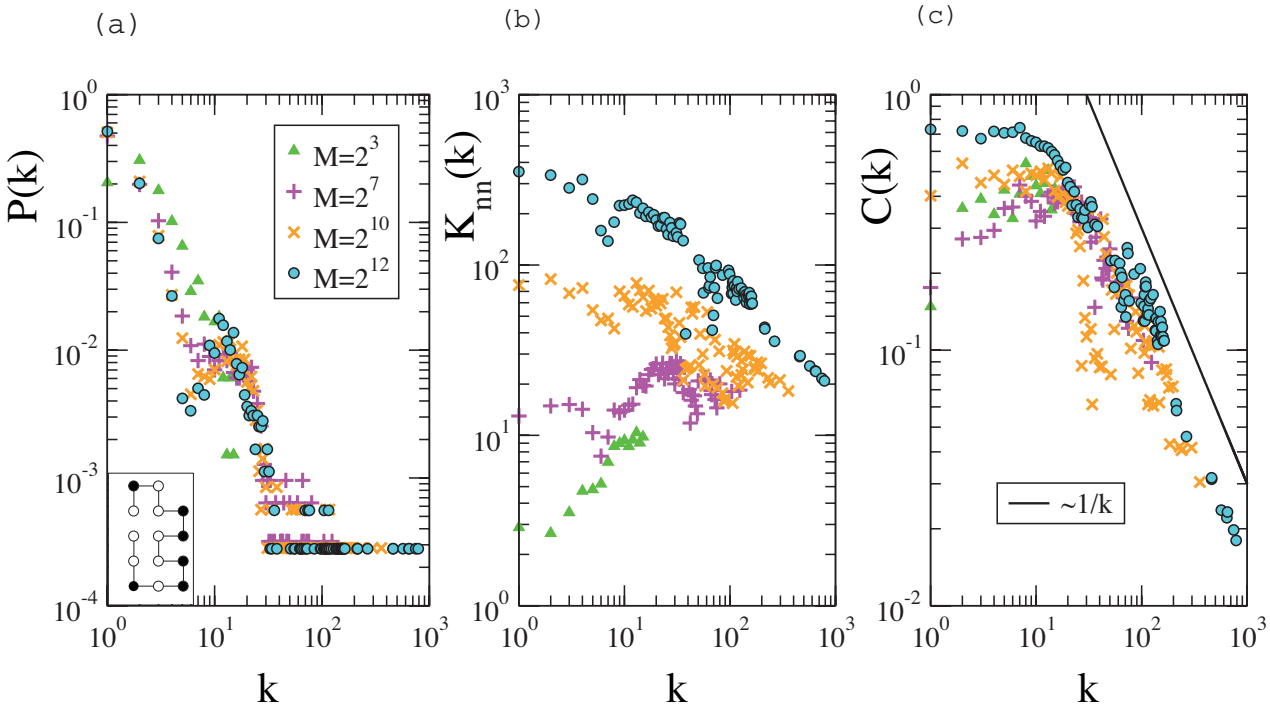

FIG. 5. (Color online) Network topology for the random lattice heteropolymer CSN at different values of the parameter M. (a) Degree distribution. The most visited configuration of the heteropolymer is displayed in the inset. (b) Average neighbor connectivity. (c) Clustering coefficient.

configuration is merely minus the number of its H-H contacts on the lattice. From a physical point of view, the cornerstone of the simulation is the appropriate adjustment of T in order to achieve an effective sampling of the lowest-energy configurations. This has been accomplished by sampling at a temperature significantly smaller than the coil-to-globule transition temperature of the polymer (i.e., $k_B T_{samp}=0.3$ and $k_B T_{trans}\approx 0.5$). The transition temperature has been identified by a thorough study of the heat capacity C_V and of two topological quantities, namely, the gyration radius and the end-to-end distance. A CSN node is defined as a single lattice configuration up to a symmetry of the lattice.

In both the beta3s and heteropolymer systems, two nodes are linked if a direct transition between them (at a given M) has been observed along the simulation. The topology of the two CSNs shows several common properties. The degree distributions $P(k)$ for beta3s and the heteropolymer are shown in Figs. 4(a) and 5(a), respectively. The distributions are robust upon varying the configuration saving time (i.e., changing the value of M) and resemble a power-law $k^{-\gamma}$ for $M \geqslant 1$ with exponent γ between 1.5 and 2. This behavior is qualitatively similar to what is observed for the quadratic well, while the steeper slope may result from the higher dimension (i.e., higher number of degrees of freedom) of the energy landscapes. Interestingly, the average neighbor connectivity $K_{nn}(k)$ changes significantly for different values of M. $K_{nn}(k)$ is shown in Figs. 4(b) and 5(b). For $M<100$, this quantity grows with the degree, whereas for larger values of M, $K_{nn}(k)$ becomes a decreasing function of k. Moreover, the assortativity coefficient q changes from positive values for $M=1$ to negative values for $M>1000$, indicating an assortative-to-disassortative transition (see Fig. 2).

The clustering coefficient $C(k)$ converges toward a general decreasing behavior for large M [Figs. 4(c) and 5(c)]. In previous work, the presence of an apparent scaling in $C(k)$ had been interpreted as the signature of a hierarchical organization of the nodes in the native state of beta3s [16]. However, a comparison between the $C(k)$ of beta3s for different values of M and of the quadratic well (see Fig. 1) strongly suggests that this decay does not indicate node hierarchy as presented in [21]. First, the quadratic well underlying the CSN does not present a hierarchical organization as in [21]. Second, it should be noticed that, in the uncorrelated regime, nodes lying at the bottom of the basin are strongly connected together, giving rise to an almost complete subgraph. In this regime, nodes with low degree are unlikely to be linked together but tend to connect to high-degree nodes (bottom configurations). These two effects are indeed sufficient to explain the decay observed in $C(k)$ without invoking a node hierarchy.

For the CSN of beta3s, there is no rigorous evidence that an uncorrelated regime is reached for $M>100$. However, analysis of the transition probabilities (i.e., link weights) can account for this behavior. In Fig. 6 the relation between $\log_{10}(w_i/w_1)$ and $\log_{10}(w_{i\to 1}/w_{2\to 1})$ is shown, where w_i and $w_{i\to 1}$ indicate the weight of node i and the weight of the link between nodes i and 1, respectively. Index 1 stands for the most populated node of the network. These logarithms have a physical meaning reflecting the configuration free energy $\Delta F_i \sim -k_B T \log_{10}(w_i)$ and the free-energy barrier between different configurations $\Delta F_{i\to j} \sim -k_B T \log_{10}(w_{i\to j})$. For $M=1$, the relation between the two free energies is not linear.

In other words, nodes with similar weights might be separated from node 1 by free-energy barriers of very different size (for instance the two nodes with $\Delta F_i \approx 1$ in Fig. 6). Choosing higher M increases the correlation between node and link weights. For $M=10^5$, $\Delta F_{i\to j}$ grows linearly with ΔF_i, indicating that link weights depend only on w_i. This behavior provides strong evidence for an uncorrelated sampling.

It is essential to stress that the uncorrelated regime is a frequent scenario when dealing with long sampling MD simulations. These simulations explore transitions between several energy basins, for example, when investigating the large configurational changes characterizing the protein folding. In these cases, the configuration saving time is usually set to large values for computational reasons, resulting in an intrabasin uncorrelated regime. Finally, it is important to note that these results have been obtained for CSNs originating from a single-basin energy landscape. In the case of networks describing fully sampled landscapes presenting a large number of basins the network topology might change since it reflects the contributions from different basins.

VI. CONNECTION WITH FITNESS (HIDDEN VARIABLE) MODELS

The scaling behavior in several networks has triggered a vast effort in modeling complex networks [22]. Of particular interest for CSNs is the model based on a fitness parameter on the nodes [27–29]. In the original fitness model [27], two nodes are connected with probability 1 if the sum of their fitness exceeds a given threshold. In the case of CSNs reflecting a single enthalpic energy basin (as in this work), the fitness of a node is given by $-U(\mathbf{x})$. Equation (4) with $n=1$ shows that nodes are connected with probability 1 if the sum of their fitnesses is higher than a threshold given by $-\ln(W_0^2 a^{2D} N)$. In addition, there is also a probability to connect nodes of high energy, given by $NP(\mathbf{x}_2)P(\mathbf{x}_1)$. This formulation shows that CSNs fall in the large class of networks whose nodes are described by a fitness parameter, also referred to as a hidden variable [28]. Notably, the scaling properties of the model presented in the papers mentioned above are in good agreement with the $P(k)$, $K_{nn}(k)$, and $C(k)$ observed in the uncorrelated case.

VII. CONCLUSIONS

The scaling behavior observed in the CSN topology has been investigated in the quadratic-well model, the native state of a triple stranded β-sheet peptide, and a lattice heteropolymer model. Despite the important differences between these systems, some universality has been observed. In particular, three main results have clearly emerged. First, in the limit of very large configuration saving times (uncorrelated regime), an analytical approximation (first order) for the degree distribution, the average neighbor connectivity, and the clustering coefficient can be carried out. Comparison between the analytical predictions and the results obtained from the simulation of the dynamics in a quadratic well shows that, in the limit considered, the analytical solution

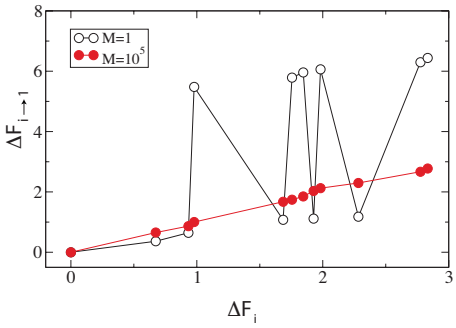

FIG. 6. (Color online) Relation between the free-energy barrier to the configuration at the bottom of the native state basin ($\Delta F_{i\to 1}$), and the configuration free energy (ΔF_i) for the most visited nodes of the beta3s network. Empty and full dots represent the $M=1$ and 10^5 cases, respectively.

describes correctly the CSN topology. These results allow for the interpretation of the topology observed in complex CSNs which cannot be tackled analytically, like the ones describing the native state of a β-sheet peptide or the low-energy configurations of a lattice heteropolymer. Second, the variation of the configuration saving time induces remarkable changes in the CSN topology. For small saving times, the network exhibits an assortative regime. On the other hand, with increasing the saving time, a disassortative behavior is observed.. Third, the emergence of a decaying tail in the clustering coefficient, which had been suggested to bear the signature of a hierarchical organization of the nodes in the native state of the β-sheet peptide, is in fact a consequence of uncorrelated sampling.

ACKNOWLEDGMENTS

F.R. thanks A. Caflisch, M. Karplus, and S. Krivov for a critical reading of the manuscript. D.M.D.L. thanks Maciej Kurant for fruitful discussions. D.G. acknowledges the financially support of COSIN (Grant No. FET Open IST 2001-33555), DELIS (Grant No. FET Open 001907), and the SER-Bern (Grant No. 02.0234).

APPENDIX

For the case of the quadratic well in $D=2$, the derivation of the degree distribution [Eq. (13)] is performed by first calculating the degree of a node at distance r. If $r \leqslant \sqrt{B}$, Eq. (5) reads:

$$k(r) = \frac{2\pi}{a^2}\int_0^{\sqrt{B-r^2}} r_1 dr_1 + \frac{2\pi}{a^2}\int_{\sqrt{B-r^2}}^{\infty} r_1 dr_1 \frac{a^4 N}{(2\pi)^2} e^{-(r_1^2+r^2)/2}$$
$$= \frac{2\pi}{2a^2}(B-r^2) + \frac{a^2 N}{2\pi} e^{-B/2}$$
$$= \frac{2\pi}{2a^2}(2+B-r^2). \tag{A1}$$

If $r > \sqrt{B}$, Eq. (5) reads

$$k(r) = \frac{2\pi}{a^2}\int_0^\infty r_1 dr_1 \frac{a^4 N}{(2\pi)^2} e^{-(r_1^2+r^2)/2} = \frac{a^2 N}{2\pi} e^{-r^2/2}. \quad \text{(A2)}$$

Equation (8) is calculated using the properties of the $\delta(f(r))$ function. For a given function $f(r)$ with n simple zeros $f(r_i^*)=0$, $f'(r_i^*)\neq 0$, $i=1,\ldots,n$, it is possible to write $\delta(f(r))=\Sigma_{i=1}^n \delta(r-r_i^*)/|f'(r_i^*)|$. Hence r^* is given by inverting Eqs. (A1) and (A2)

(1) If $r<\sqrt{B} \Leftrightarrow k>2\pi/a^2$,

$$r^* = \sqrt{2\left(1+\frac{B}{2}\right) - \frac{a^2}{\pi}k} \Leftrightarrow P(k) \sim \frac{r^*}{|(2\pi/a^2)r^*|} \sim \text{const.}$$

(2) If $r<\sqrt{B} \Leftrightarrow k<2\pi/a^2$,

$$r^* = \sqrt{2\ln\frac{a^2 N}{2\pi k}} \Leftrightarrow P(k) \sim \frac{r^*}{(a^2N/2\pi)r^* e^{-(r^*)^2/2}} \sim \frac{1}{k}$$

The average neighbor connectivity is calculated using Eq. (9):

$$K_{nn}(k) = \begin{cases} \dfrac{2\pi^2}{a^4}2\dfrac{1}{k} + \dfrac{2\pi}{a^2}\left(1+\dfrac{B^2}{2}\right) - \dfrac{1}{2}k + \dfrac{\pi}{a^2}e^{-1-B/2}\dfrac{1}{k}e^{a^2/2\pi k} & \text{if } k > \dfrac{2\pi}{a^2}, \\ B\dfrac{\pi}{a^2} + \dfrac{4\pi^3}{a^6 N} & \text{if } k \leqslant \dfrac{2\pi}{a^2}. \end{cases} \quad \text{(A3)}$$

The expression for the clustering coefficient is slightly more complex since it requires us to distinguish between several cases according to the possible values of r. Of particular interest is the case of large k (i.e., small r). If $r<\sqrt{B/2} \Leftrightarrow k>(2\pi/a^2)(1+B/4)$, solving the integral of Eq. (7) gives

$$C(k) = \frac{N^{\triangle}(k)}{k^2} = \frac{(2\pi)^2}{a^4}\frac{1}{k^2}\left[\frac{1}{2}B + \frac{B^2}{8} - 1 - \left(\frac{a^2}{2\pi}k - 1 - \frac{B}{2}\right)^2\right] + \frac{(2\pi)^2}{a^4}\frac{1}{k^2}\left[\exp\left(\frac{a^2}{2\pi}k - 1 - \frac{B}{2}\right) + \frac{\pi^2}{a^4}\exp\left(-\frac{a^2}{\pi}k + 2 + B\right)\right]. \quad \text{(A4)}$$

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[3] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[5] G. Caldarelli, *Scale-Free Networks* (Oxford University Press, Oxford, 2007).
[6] G. Caldarelli, A. Capocci, and D. Garlaschelli, e-print arXiv:cond-mat/0611201.
[7] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[8] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, Comput. Phys. Commun. **91**, 43 (1995).
[9] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. R. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman, Comput. Phys. Commun. **91**, 1 (1995).
[10] R. Ru, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[11] A. Scala, L. A. N. Amaral, and M. Barthélémy, Europhys. Lett. **55**, 594 (2001).
[12] J. P. K. Doye, Phys. Rev. Lett. **88**, 238701 (2002).
[13] J. P. K. Doye and C. P. Massen, J. Chem. Phys. **122**, 084105 (2005).
[14] S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[15] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **101**, 14766 (2004).
[16] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[17] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao, Proc. Natl. Acad. Sci. U.S.A. **104**, 1817 (2007).
[18] A.-L. Barabási and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[19] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[20] D. Gibson, J. Kleinberg, and P. Raghavan, *HYPERTEXT '98: Proceedings of the 9th ACM Conference* (ACM Press, New York, 1998).
[21] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[22] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[23] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[24] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[25] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).
[26] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[27] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[28] M. Boguna and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).
[29] V. D. P. Servedio, G. Caldarelli, and P. Buttà, Phys. Rev. E **70**, 056126 (2004).
[30] W.-X. Wang, B. Hu, B.-H. Wang, and G. Yan, Phys. Rev. E **73**,

016133 (2006).
[31] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct., Funct., Genet. **46**, 24 (2002).
[32] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10780 (2000).
[33] A. Cavalli, P. Ferrara, and A. Caflisch, Proteins: Struct., Funct., Genet. **47**, 305 (2002).
[34] J. Gsponer, U. Haberthür, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **100**, 5154 (2003).
[35] P. Carter, C. A. F. Andersen, and B. Rost, Nucleic Acids Res. **31**, 3293 (2003).
[36] K. Dill, Biochemistry **24**, 1501 (1985).
[37] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[38] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[39] M. Karplus and A. Sali, Curr. Opin. Struct. Biol. **5**, 58 (1995).
[40] D. Thirumalai and S. A. Woodson, Acc. Chem. Res. **29**, 433 (1996).
[41] A. D. Sokal, Nucl. Phys. B, Proc. Suppl. **47**, 172 (1996).
[42] N. Madras and G. Slade, *The Self-Avoiding Walk* (Birkhauser, Boston, 1996).