Stochastic dynamics of model proteins on a directed graph

Lorenzo Bongini,[1,2,*] Lapo Casetti,[1,2,3,†] Roberto Livi,[1,2,3,‡] Antonio Politi,[2,4,§] and Alessandro Torcini[2,3,4,∥]
[1]*Dipartimento di Fisica, Università di Firenze, via Sansone 1, 50019 Sesto Fiorentino, Italy*
[2]*Centro Interdipartimentale per lo Studio delle Dinamiche Complesse, via Sansone 1, 50019 Sesto Fiorentino, Italy*
[3]*INFN–Sezione di Firenze, via G. Sansone 1, 50019 Sesto Fiorentino, Italy*
[4]*Istituto dei Sistemi Complessi, CNR, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Italy*

(Received 14 January 2009; published 30 June 2009)

A method for reconstructing the potential energy landscape of simple polypeptidic chains is described. We show how to obtain a faithful representation of the energy landscape in terms of a suitable directed graph. Topological and dynamical indicators of the graph are shown to yield an effective estimate of the time scales associated with both folding and equilibration processes. This conclusion is drawn by comparing molecular dynamics simulations at constant temperature with the dynamics on the graph, defined as a temperature-dependent Markov process. The main advantage of the graph representation is that its dynamics can be naturally renormalized by collecting nodes into "hubs" while redefining their connectivity. We show that the dynamical properties at large time scales are preserved by the renormalization procedure. Moreover, we obtain clear indications that the heteropolymers exhibit common topological properties, at variance with the homopolymer, whose peculiar graph structure stems from its spatial homogeneity. In order to distinguish between "fast" and "slow" folders, one has to look at the kinetic properties of the corresponding directed graphs. In particular, we find that the average time needed to the fast folder for reaching its native configuration is two orders of magnitude smaller than its equilibration time while for the bad folder these time scales are comparable.

DOI: 10.1103/PhysRevE.79.061925 PACS number(s): 87.15.A−, 87.15.hm, 82.35.Lr, 05.40.−a

I. INTRODUCTION

Numerical simulations are quite often an effective approach for studying dynamical properties of systems with many degrees of freedom. In particular, when the forces acting among the constituent particles (atoms, molecules, monomers, etc.) can be well approximated as pairwise interactions, molecular dynamics (MD) is a useful tool for investigating complex dynamical phenomena.

The main practical limitation to this basic approach is the existence of relaxation processes, whose time scales are several orders of magnitude longer than the typical time scale of the microscopic dynamics. Models of structural glasses exhibit such puzzling features that are associated with the anomalously high viscosity of these amorphous materials and with the aging phenomenon [1]. Something similar occurs in proteins, where the folding or the equilibration processes may proceed over exceedingly long time scales with respect to the microscopic ones [2]. In all of these cases, reproducing the interesting phenomena starting from a microscopic description can be very expensive in terms of CPU time and mass storage memory. A possible solution to such difficulties could arise from a drastic simplification of the microscopic model by coarse graining it to derive an effective representation in terms of macroscopic modes. Such a strategy is normally implemented to obtain a hydrodynamic description after having removed the fast time scales. This results can be achieved by the so-called "projection technique," which is at the basis of the linear-response theory by Green and Kubo [3]. Unfortunately, it applies successfully only to simple gases and liquids [4].

More effective strategies to characterize anomalously long relaxation processes should exploit the peculiar structure of the energy landscape [5,6]. In particular, the phase space is often hierarchically partitioned into loosely connected regions: a trajectory may wander over a very long time inside a restricted area before finding a way out through some "bottleneck" and enter unexplored regions. For temperatures small enough with respect to the energy barriers of the bottlenecks, the situation may look quite similar to the phenomenon of ergodicity breaking that characterizes phase transitions in statistical mechanics. On the other hand, since in MD the number of degrees of freedom is finite, there are no true phase transitions and the whole phase space can be eventually explored (with the exception of exceedingly small energies).

On a relatively fine scale, the potential energy landscape (PEL) can be seen as a collection of basins of attraction of the local minima of the potential energy [5,7,8]. The basin of attraction of a local minimum is the set of points in phase space whose gradient dynamics converges to that minimum. Adjacent basins of attractions are separated by the stable manifolds of other stationary points, the saddles. By associating a different symbol to each basin of attraction, a trajectory can be encoded as a sequence of symbols and of the corresponding residence times. Accordingly, MD can be effectively replaced by a stochastic dynamics defined on a connected graph; an interesting application of this approach has been already reported for polyalanine [9]. The local minima, or, equivalently, their basins of attraction, are the nodes of

[*] bongini@fi.infn.it
[†] casetti@fi.infn.it
[‡] livi@fi.infn.it
[§] antonio.politi@isc.cnr.it
[∥] alessandro.torcini@isc.cnr.it

this graph. Two nodes are connected by a link if their basins of attractions are adjacent, in which case two (generally different) transition rates can be defined. We estimate the rates by exploiting a suitable high-dimensional generalization of the Arrhenius law (the so-called Langer's formula [10,11]; see Sec. II C). As expected, the validity of this expression depends crucially on the temperature, and in particular it works for temperatures that are comparable to or lower than the typical energy barriers; otherwise the memory of previous jumps cannot be neglected.

In this paper we test this approach by studying simple model proteins whose dynamical features were analyzed in previous publications [12,13]. In Sec. II, we summarize the protein model and the method used for reconstructing minima and saddles of the PEL. We also show that the strategy proposed in [13] can be improved by adopting a suitable search for identifying shortcuts that connect minima separated by large conformational distances. Later in this section we show how one can plug a Markov-chain structure onto the reconstructed directed graph. In Sec. III we discuss how such a representation can be used to determine equilibrium and nonequilibrium properties, to be compared with those obtained by MD simulations. As expected, the "graph" approximation is effective in a range of temperatures close to the folding one. In Sec. IV we comment about the advantages of the graph approximation with respect to MD simulations. In fact, one can easily realize that a suitable reconstruction of the PEL, including a sufficient sampling of minima and saddles, requires a considerable numerical effort, comparable with MD simulations. However, these efforts are made once and for all and do not need to be repeated for different temperatures. Moreover, the graph dynamics (GD) can be further simplified without loosing the essential details. This can be done by means of a renormalization procedure that amounts to progressively removing irrelevant nodes and saddles. Furthermore, additional information about the nature of model proteins can be obtained by comparing the static properties of the graphs with those of random directed graphs. This allows us to conclude that some general static features are common to any model of a polypeptidic chain. The main specific signatures of a protein specimen (*fast folder*) seem rather to be associated with its dynamical features. This is not completely unexpected, although quite often one finds in the literature claims about specific static properties of the energy landscape as intrinsic to real proteins [14]. Our analysis at least challenges this widespread belief.

II. MODEL, ITS ENERGY LANDSCAPE, AND THE DIRECTED GRAPH

A. Simple toy model of polypeptidic chains in two dimensions

For the sake of simplicity, we use a simple toy model to test the idea of approximating the thermalized dynamics of a polypeptidic chain with a stochastic dynamics on a directed graph. The model, introduced in [15], is a slight modification of the 2*d* off-lattice HP model originally proposed by Stillinger *et al.* in [16]. It is defined by the Hamiltonian

$$H = T + V, \tag{1}$$

where

$$T = \sum_{i=1}^{L} \frac{p_{x,i}^2 + p_{y,i}^2}{2} \tag{2}$$

is the kinetic energy, while the intramolecular potential

$$V = \sum_{i=1}^{L-1} V_1(r_{i,i+1}) + \sum_{i=2}^{L-1} V_2(\theta_i) + \sum_{i=1}^{L-2} \sum_{j=i+2}^{L} V_3(r_{ij}, \xi_i, \xi_j) \tag{3}$$

is composed of three terms: a stiff nearest-neighbor harmonic potential V_1, which keeps the bond distance almost constant, a three-body potential V_2, which measures the energetic cost of local bending, and a Lennard-Jones potential V_3 acting between all pairs of monomers i and j such that $|i-j|>1$,

$$V_1(r_{i,i+1}) = \alpha(r_{i,i+1} - r_0)^2,$$

$$V_2(\theta_i) = \frac{1 - \cos\,\theta_i}{16},$$

$$V_3(r_{i,j}) = \frac{1}{r_{i,j}^{12}} - \frac{c_{i,j}}{r_{i,j}^{6}}. \tag{4}$$

This Hamiltonian schematizes a real protein as a one-dimensional chain of L pointlike monomers of two types, hydrophobic (H) and polar (P). Accordingly, a heteropolymer is identified by a sequence of binary symbols indicating the monomer type. The monomers are assumed to have the same unitary mass (all parameters are expressed in terms of adimensional arbitrary units). The space coordinates of the ith monomer are $\mathbf{q}_i=(x_i,y_i)$ and their conjugated momenta are $\mathbf{p}_i=(p_{x,i},p_{y,i})=(\dot{x}_i,\dot{y}_i)$. The variable $r_{i,j}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$ is the distance between ith and jth monomers and θ_i is the bond angle at the ith monomer. The parameters α and r_0 are fixed to the values 20 and 1, respectively. V_3 is the only potential term that depends on the nature of the monomers. In fact, $c_{i,j}=\frac{1}{8}(1+\xi_i+\xi_j+5\xi_i\xi_j)$, where $\xi_i=1(\xi_i=-1)$ if the monomer is hydrophobic (polar). As a result, the interaction is attractive if both residues are either hydrophobic or polar (with $c_{i,j}=1$ and 1/2, respectively), while it is repulsive if the residues belong to different species ($c_{ij}=-1/2$).

Here, we focus our investigation on three sequences of 20 monomers that represent the three classes of different folding behaviors observed in this model:

(i) [S0] is a homopolymer composed of 20 H residues;

(ii) [S1]=[*HHHP HHHP HHHP PHHP PHHH*] is a sequence that has been identified as a fast folder in [17]; and

(iii) [S2]=[*PPPH HPHH HHHH HHHP HHPH*] is a randomly generated sequence that has been identified as a slow folder in [15].

The three characteristic temperatures T_θ, T_f, and T_g of each sequence have been determined in [12] by MD simulations, where the chains have been put in contact with a Langevin heat reservoir. Their numerical values are reported in Table I together with the number n_0 of minima directly connected to the native state via a first-order saddle. Here n_0

TABLE I. The collapse transition temperature T_θ, the folding temperature T_f, the "glassy" temperature T_g, and the number n_0 of minima directly connected to the native configuration for the sequences S0 (homopolymer), S1 (fast folder), and S2 (slow folder).

	S0	S1	S2
T_θ	0.16	0.11	0.13
T_f	0.044	0.061	0.044
T_g	0.022	0.048	0.025
n_0	57	66	69

has been obtained by a more refined analysis with respect to that performed in [13].

Several distances can be defined in order to distinguish between two configurations $\mathcal{C}_1$ and $\mathcal{C}_2$ of a two-dimensional chain. A particularly simple one is the angular distance

$$d_\theta(\mathcal{C}_1,\mathcal{C}_2) = \frac{1}{L-2}\sum_{n=1}^{L-2} |\theta_i(\mathcal{C}_1) - \theta_i(\mathcal{C}_2)|, \qquad (5)$$

where

$$\theta_i(\mathcal{C}) = \frac{(\mathbf{q}_i - \mathbf{q}_{i+1}) \cdot (\mathbf{q}_{i+1} - \mathbf{q}_{i+2})}{r_{i,i+1} r_{i+1,i+2}} \qquad (6)$$

is the ith backbone angle of configuration $\mathcal{C}$ represented by the coordinates $\mathbf{q}_i$.

B. Reconstruction of the energy landscape

The PEL can be reconstructed by implementing an effective strategy to identify the local minima of the potential energy and the first-order saddles (i.e., extremal points that are local maxima only along one direction) separating them in the configuration space. A first extensive search for minima can be performed by sampling a sufficiently large number of MD trajectories at a temperature T close to T_f, as we wish to explore the relevant processes connected with the protein (un)folding. The constant temperature constraint is imposed by attaching a Langevin heat bath to all degrees of freedom. The strength of the coupling with the heat bath is given by the dissipation rate $\gamma=7$: this value, expressed in the adimensional units of the model, has been estimated from the knowledge of the relaxation rate of an amino acid in a solvent (typically, water) [12]. Trajectories are sampled at a time pace $\Delta t=0.1$ (time will be always expressed in the natural adimensional units), which is a bit longer than the residence time within a typical basin of attraction. The basin of attraction is identified by taking each sampled configuration as the initial condition for an overdamped dynamics (see [13]), converging to the corresponding minimum of the potential energy. As shown in [13], a large number of minima of the PEL of sequences S0, S1, and S2 can be identified by generating $O(10^3)$ Langevin trajectories of duration $t=10^3$ in the natural time units of model (1).

Once this preliminary set of minima has been constructed, one is interested in determining pairs of adjacent minima, i.e., minima that are connected through a first-order saddle (this is the typical case for smooth potentials such as those invoked in Hamiltonian (1)—see [18] for the specific requirements concerning the smoothness of the potential). In the last decade, various methods have been proposed to design efficient algorithms for the search of saddles [19–21]. Unfortunately, none of these approaches is so effective to be really useful in this context. As shown in [13], one can take advantage of a metric criterion to identify pairs of potentially adjacent minima: they are typically separated by an angular distance d_θ smaller than $d_\theta^{thr}=0.2$ (see [13]). However, in spite of the effectiveness of this criterion, one cannot expect to identify all the relevant saddles involved in the folding process. Actually, it has been observed that the PELs of sequences S1 and S2 contain a relatively small set of first-order saddles connecting pairs of minima, whose d_θ is definitely larger than d_θ^{thr}. In order not to miss such saddles, one can use a more refined strategy (see [13]). The identified minima are taken as initial conditions of the Langevin dynamics at $T=T_f$. The dynamics is then let evolve until it enters the basin of attraction of a different minimum. The test is made by determining every $\Delta\tau=10^{-3}$ units the corresponding basin by using the sampled configurations as initial conditions for an overdamped dynamics [22]. Whenever new minima are discovered, they are added to the database. Finally, after refining the position of the corresponding saddle with the procedure described in [13], the saddle itself is added to the database of the links between adjacent minima.

TABLE II. Number of minima, N, and saddles, S, for the three analyzed sequences. The number of minima, N_f, and saddles, S_f, below the folding energy E_f is also reported.

	S0	S1	S2
N	180156	87580	110524
S	349197	213219	304303
N_f	99797	17726	35852
S_f	276958	85014	150809

A number of minima N and saddles S obtained for each sequence are reported in the first two rows of Table II. There, we have also reported the number of minima (N_f) and saddles (S_f), with a potential energy smaller than the "folding energy" $E_f=V_0+k_BT_f(2L-3)/2$. We expect that these subsets should contain the main elements associated with the folding process. Notice that the fraction of minima and saddles below E_f reduces much more for the heteropolymer sequences (S1 and S2) than for the homopolymer (S0). This is a first indication of quantitative differences among the PELs of different sequences.

C. Directed graph

The minima and first-order saddle database can be used to construct a directed graph. Each node of the graph corresponds to one minimum (or, equivalently, to its basin of attraction). In what follows we assume that the N nodes are ordered in increasing values of their potential energy and assign them an index i that runs from 1 to N. Accordingly,

$i=1$ corresponds to the so-called native state. A link between node i and node j means that they are connected by a first-order saddle $s_{i,j}$. The transition rates $\Gamma_{i,j}$ and $\Gamma_{j,i}$ are thereby associated to the link by means of Langer's formula [10,11], which generalizes the usual Arrhenius formula by including the entropic factors that are estimated from the curvatures of minima and saddles of the potential energy hypersurface. The expression for $\Gamma_{i,j}$ reads

$$\Gamma_{i,j} = \frac{\omega_{\parallel\ i,j}}{\pi\gamma}\frac{\prod_{k=1}^{L'}\omega_i^{(k)}}{\prod_{k=1}^{L'-1}\omega_{\perp\ i,j}^{(k)}}\exp\left(-\frac{V(s_{i,j})-V(i)}{k_BT}\right), \tag{7}$$

where $\{\omega_i^{(k)}\}$ are the $L'=2L-3$ nonzero eigenfrequencies of the minimum associated with node i, $\{\omega_{\perp\ i,j}^{(k)}\}$ are the $L'-1$ nonzero frequencies corresponding to the contracting directions of $s_{i,j}$, and $\omega_{\parallel\ i,j}$ is the frequency associated with the only expanding direction of $s_{i,j}$. The dissipation rate γ is the same, used for defining the Langevin dynamics mentioned in Sec. II B. The exponential factor in Eq. (7) depends on the height of the energy barrier $V(s_{i,j})-V(i)$, where we have used the short-hand notations $V(s_{i,j})$ and $V(i)$ to denote the values of the potential energies of the saddle $s_{i,j}$ and of the minimum i, respectively. Finally, the expression for $\Gamma_{j,i}$ is obtained by exchanging the index i with j and noticing that $\omega_\perp$ and $\omega_\parallel$ are both symmetric (the same saddle contributes to $\Gamma_{i,j}$ and $\Gamma_{i,j}$).

The nonsymmetric $N\times N$ connectivity matrix Γ provides a faithful description of the original dynamics as long as one can neglect the memory of previous transitions. This is made explicit in the master equation ruling the evolution of the probability $P_i(t)$ that the polymer is in node i at time t,

$$\frac{dP_i(t)}{dt} = \sum_{j=1}^{N} P_j(t)\Gamma_{j,i} - P_i(t)\sum_{j=1}^{N}\Gamma_{i,j}. \tag{8}$$

This master equation can be cast into matrix form,

$$\frac{dP(t)}{dt} = -WP(t), \tag{9}$$

where $P(t)$ is a vector of dimension N at time t, whose elements are $P_i(t)$, while the entries of the Laplacian matrix W are given by the expression

$$W_{i,j} = \delta_{i,j}\sum_{k=1}^{N}\Gamma_{j,k} - \Gamma_{j,i}. \tag{10}$$

W is a nonsymmetric real matrix with positive diagonal elements and whose rows and columns sum up to zero: according to Gershgorin's theorem [23], all its eigenvalues r_i, $i=1,\ldots,N$, are real and positive apart from the null eigenvalue $r_1=0$ (usually, r_i's are ordered in increasing value of the index i). The corresponding eigenvectors are denoted with $w^{(i)}$. The stationary probability coincides with the first eigenvector $w^{(1)}$. Its components are

$$w_i^{(1)} = \alpha\frac{e^{-[V(i)/k_BT]}}{\prod_{k=1}^{L'}\omega_i^{(k)}}, \tag{11}$$

where α is a suitable normalization constant, such that $\Sigma_{i=1}^{N} w_i^{(1)}=1$. By combining Eqs. (11) and (7), one can verify that detailed balance is satisfied, namely, that

$$w_i^{(1)}\Gamma_{i,j} = w_j^{(1)}\Gamma_{j,i}. \tag{12}$$

In general, the nonzero eigenvalues of W, r_i with $i=2,\ldots,N$ can be determined only numerically; detailed balance simplifies the calculations as it makes possible to transform W into a symmetric matrix $\mathcal{W}=T^{-1}WT$ [24], with

$$T = \begin{pmatrix} \sqrt{w_1^{(1)}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{w_N^{(1)}} \end{pmatrix}. \tag{13}$$

By expanding the initial probability distribution in terms of the eigenvectors, $P(0)=\Sigma_{k=1}^{N} c_k w^{(k)}$ [with real c_k and the normalization condition $c_1=1$ (see Appendix A)] and the probability $P_i(t)$ to be on the node i at time t can be written as

$$P_i(t) = \sum_{k=1}^{N} c_k w_i^{(k)} e^{-r_k t}. \tag{14}$$

This expression stems from the orthogonality of the eigenvectors $w^{(k)}$ (see Appendix A).

Another consequence of Eq. (12) is that the stationary probability flux does depend neither on the starting nor on the arrival minima. In the harmonic approximation, it depends only on the energy and on the curvature of $s_{i,j}$,

$$J_{i,j} = w_i^{(1)}\Gamma_{i,j} = \frac{\omega_{\parallel\ i,j}}{\prod_{k=1}^{L'-1}\omega_{\perp\ i,j}^{(k)}} e^{[-V(s_{i,j})/k_BT]}. \tag{15}$$

We conclude this section by observing that the topological properties of the directed graph can be studied by introducing the topological connectivity matrix Γ_0, where all the nonzero elements of Γ are set to 1. By replacing Γ_0 with Γ in Eq. (10), one obtains the topological Laplacian matrix W_0. We show in the following sections that the knowledge of the latter matrix allows us to infer some general properties of the corresponding graph. For instance, the power-law behavior of the low-frequency component of its spectral density determines the spectral dimension of the graph [25]. This extends the concept of Euclidean dimension to graphs that are not defined on a regular lattice.

III. COMPARISON BETWEEN MD AND THE MARKOV CHAIN ON THE DIRECTED GRAPH

In this section we investigate to what extent the stochastic dynamics defined by the Laplacian matrix W is consistent with MD simulations at least for a temperature T close to T_f. A first simple test can be performed on the expectation values of equilibrium properties. These can be analytically de-

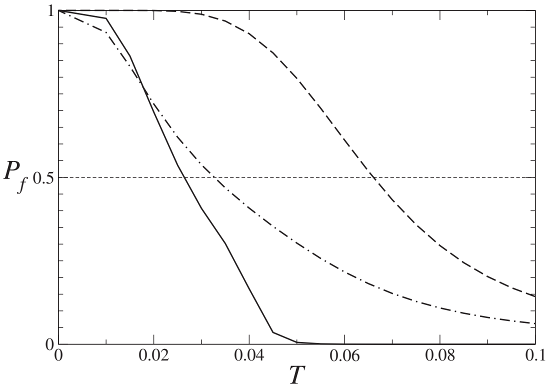

FIG. 1. Folding probability P_f estimated from Eq. (11) as a function of temperature T. The full dotted-dashed and dashed lines correspond to S0, S2, and S1, respectively.

termined on the graph from the stationary probability vector $w^{(1)}$ defined in Eq. (11). Since equilibrium properties depend only on the identified minima, this test can provide a quantitative verification of the reliability of the algorithm used for locating them in the PEL. In particular, we have compared MD estimates of the folding temperatures T_f of the considered sequences with the same quantities computed from the probability distribution, as given by the eigenvector $w^{(1)}$. In Fig. 1 we plot the equilibrium probability P_f that a sequence is in the native state or in the set of minima directly connected with it as a function of temperature T. In practice this amounts to measuring the fraction of "folded" sequences at a given T. Then, we apply the same criterion used in equilibrium MD simulations [13]: T_f is determined as the temperature value at which 50% of the polymer configurations are in the folded state. The results obtained with MD and with the directed graph representation are reported in Table III. There is a reasonable agreement for the good folder S1, while for both S0 and S2, T_f is underestimated. This suggests that we are missing several paths that allow returning to the native state: it is not surprising that this limitation manifests itself precisely in the cases of the most glassy landscapes.

Next, we have performed a more accurate test by determining the average exit time from a given region in the PEL and the first passage time t_f from the native state. The former aims at verifying the conjecture that MD corresponds to a sequence of thermally activated transitions through the nodes of the directed graph; the latter allows us to estimate the time scales involved in the folding process.

TABLE III. Folding temperatures as computed by means of MD simulations (Table I) and of the analytical expression of the stationary distribution on the directed graph.

	MD	Graph
S0	0.044	0.026
S1	0.061	0.066
S2	0.044	0.033

TABLE IV. Comparison of average escape times from three different sets of initial conditions computed by MD and GD simulations at temperatures $T=0.04$ and $T=0.06$ for the sequence S1. The first shell contains n_0 minima, while the set $\mathcal{M}$ contains 2341 minima.

	From the native state $T=0.04$	 $T=0.06$
MD	4.1×10^3	1.6×10^2
GD	3.5×10^3	85
	From the first shell $T=0.04$	 $T=0.06$
MD	2.3×10^5	4.4×10^2
GD	1.3×10^5	2.4×10^2
	From the set $\mathcal{M}$ $T=0.04$	 $T=0.06$
MD	3.3×10^6	2×10^4
GD	1.1×10^6	7×10^2

In the GD, the average time spent at node i can be estimated as

$$\langle t_i\rangle = \frac{1}{\sum_k \Gamma_{i,k}}, \tag{16}$$

where the index k runs over all the nodes directly connected with i. Moreover, the probability of moving from node i to node j is

$$\Pi_{i,j} = \Gamma_{i,j}\langle t_i\rangle. \tag{17}$$

Accordingly, a trajectory on the directed graph can be represented by an ordered set of symbols $(i_1, i_2, i_3, \ldots, i_n)$ labeling the visited nodes, while its time duration is

$$t = \sum_{\alpha=i}^{n} \langle t_\alpha\rangle. \tag{18}$$

Notice that the inverse dissipation rate $1/\gamma$ is the time scale that allows establishing a correspondence between MD and GD [see Sec. II B and Eq. (7)].

We have determined the average exit time of the sequence S1 from (i) the native state, (ii) the first shell (i.e., the set of the n_0 minima directly connected with the native one), and (iii) the set of minima $\mathcal{M}$, whose angular distance d_θ from the native state is smaller than 0.4 (this latter set contains 2341 minima, including the first shell). MD averages have been performed over 10^3 trajectories, starting from each minimum in the considered set, while GD averages have been performed over 10^4 stochastic paths. The results for $T=T_g=0.4$ and $T=T_f=0.6$ are reported in Table IV. There one can see that the agreement progressively deteriorates for initial regions of larger sizes: GD increasingly underestimates the escape times. This is quite surprising, as one would have *a priori* expected that GD escape times become longer because of undetected escape routes. The only explanation for

TABLE V. Comparison of average first passage times at the native state from three different sets of initial conditions computed by MD and GD simulations at temperatures of $T=0.04$ and $T=0.06$ for the sequence S1. The number of minima in the first shell and that in the set $\mathcal{M}$ have been reported in Sec. IV, while the set $\mathcal{N}$ contains 17 726 minima.

	From the first shell	
	$T=0.04$	$T=0.06$
MD	5.7×10^4	1.2×10^4
GD	4.3×10^4	4.6×10^3
	From the set $\mathcal{M}$	
	$T=0.04$	$T=0.06$
MD	8.5×10^3	4.1×10^3
GD	4.4×10^3	5.1×10^2
	From the set $\mathcal{N}$	
	$T=0.04$	$T=0.06$
MD	6.9×10^4	5.8×10^4
GD	4.6×10^4	5.8×10^3

the discrepancy we are able to propose is that the known overestimation of the transition rate in Langer's formula [13] becomes increasingly severe for the minima that are not too close to the native state.

We have also estimated the first passage time t_f of S1 to its native minimum for three different classes of initial conditions: the minima in the first shell, the set $\mathcal{M}$, and the yet larger set of minima $\mathcal{N}$, whose potential energy is smaller than $E_f=-3.45$ (17 726 minima). The simulations have been performed again by averaging over 10^3 MD trajectories and 10^4 GD paths. The data reported in Table V reveal a reasonably good agreement for $T=0.04$, while for $T=0.06$ the first passage time on the graph is much smaller than the MD estimate. Apart from the approximation of Langer's formula, the main reason for this discrepancy is the poor reconstruction of the PEL for high values of the potential energy. In fact, we have checked that the agreement between the two estimates improves significantly—up to a factor of 5—if we exclude, both in the GD and MD cases, all the trajectories that escape from the chosen set of initial minima. This implies that MD visits regions of the PEL that are poorly reproduces by the GD. Only for temperatures smaller than T_f, this effect is sufficiently negligible. Unfortunately, a finer sampling of the PEL would require significantly larger computational efforts even to produce small improvements.

For what concerns S2, measurements of the average folding time at $T=0.06$ starting from the set of minima with potential energy below E_f give the quantitatively consistent estimates of 1.3×10^6 for MD and 9.8×10^5 for GD. At $T=0.04$ MD simulations are practically unfeasible. In fact, GD simulations predict an average folding time $O(10^8)$, i.e., three orders of magnitude larger than for S1. The situation is even worse for MD simulations of S0. GD simulations suggest that already at $T=0.06$ the average folding time raises up to $O(10^9)$. Therefore, reliable estimates of the average folding time are possible only for $T>T_\theta$, where we know that the graph representation is unreliable [12].

IV. RENORMALIZATION OF THE DIRECTED GRAPH

Although a considerably high computational price has to be paid to eventually assemble a sufficiently accurate graph, GD has several advantages. First of all, there are no low-temperature limitations, as one has just to draw at random escape times from the single minima. Moreover, the graph structure can be easily manipulated to identify general properties. In particular, one can progressively renormalize the graph by gluing together those nodes that are more "tightly connected." More precisely, our approach consists in the following steps:

(i) Given any pair of connected nodes i and j, we compute the height of the energy barrier as

$$\sigma_m = V(s_{i,j}) - V_M, \tag{19}$$

where $V_M=\max[V(j),V(i)]$.

(ii) The energy barriers are ordered from the minimum σ_1 to the maximum value.

(iii) The two nodes i and j bridged by the smallest barrier are identified with one another. In practice, node j is eliminated, while the transition rates of node i are rescaled in such a way that the equilibrium eigenvector $\widetilde{w}^{(1)}$ reads as

$$\widetilde{w}_i^{(1)} = w_i^{(1)} + w_j^{(1)} \tag{20}$$

while the transition rates $\widetilde{\Gamma}_{k,j}$ become

$$\begin{aligned}\widetilde{\Gamma}_{k,i} &= \Gamma_{k,i} + \Gamma_{k,j},\\ \widetilde{\Gamma}_{i,k} &= \frac{\Gamma_{j,k}w_j^{(1)} + \Gamma_{i,k}w_i^{(1)}}{w_j^{(1)} + w_i^{(1)}}.\end{aligned} \tag{21}$$

(iv) In the list of barriers connecting neighboring minima, the index j is replaced everywhere with i.

(v) The two minima corresponding to the next lowest barrier are identified and the same procedure described in the previous two steps is repeated until $\sigma_m \le k_B T$.

One can easily verify that $\widetilde{w}^{(1)}$ is the equilibrium eigenvector of the corresponding renormalized evolution matrix $\widetilde{W}$. Analogously, it can be seen that the rescaled dynamical rules still satisfy the detailed balance condition (12). Notice that a similar methodology, termed "disconnectivity graph" approach, has been developed in recent years and successfully applied to both the reconstruction of the PEL [26] and of the free-energy landscape [27] of model proteins and peptides.

The renormalization procedure transforms also the topological connectivity matrix Γ_0 and the corresponding Laplacian matrix W_0 to $\widetilde{\Gamma}_0$ and $\widetilde{W}_0$, respectively. It is worth noting that the renormalization method allows, in principle, us to derive the topology of the approximating graph and the transition rates that govern the master equation at any desired temperature. The effects of temperature on the dynamics of the system can therefore be analyzed with relative computa-

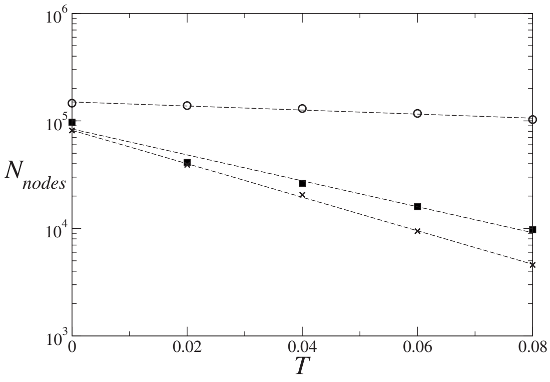

FIG. 2. (a) Number of nodes N and (b) average connectivity $\bar{\sigma}$ of the renormalized graph versus temperature T for S0 (empty circles), S1 (crosses), and S2 (filled squares). In (a) the dashed lines are exponential fits of the data, due to the adopted log-linear representation. The full lines in (b) are drawn to guide the eyes.

tional ease at variance with previous approaches based on a dynamic sampling of the conformational space [28], which require a distinct simulation effort for every given temperature. In Sec. IV A we first discuss the dependence of topological properties of the renormalized directed graph on temperature. The corresponding dynamical features will be then analyzed.

A. Topological properties of the renormalized graph

Here we investigate the dependence of the topological properties of the renormalized graph on the temperature. In Fig. 2 we show how the number of effective nodes N and the number of connections per node $\bar{\sigma}=S/N$ change with the temperature for the three sequences defined in Sec. II. Temperatures are varied in the range $0\leq T\leq 0.08$, thus encompassing both the glassy and the folding temperatures of all sequences (see Table I). In the homopolymer S0, both N and σ exhibit a very weak dependence on T (slowly decreasing and increasing, respectively). This indicates that in the explored temperature range, the graph of S0 is poorly affected by renormalization. This is due to the peculiar structure of its PEL: most minima are separated by large barriers (on this temperature scale). Conversely, in S1 and S2, N decreases by more than one order of magnitude while σ drops by a factor 2, becoming equal to the value found in S0. A simple argument can help us to understand this phenomenon. Let us assume that nodes j and i (with connectivities s_j, s_i, respectively) are assimilated and let σ denote the average connectivity before a renormalization. The contribution of the pair of nodes i and j to the deviation from the average connectivity is $\Delta=s_i+s_j-2\sigma$. After the renormalization, the contribution of the single node i is $\Delta'=s_i+s_j-c-2-\sigma$, where c is the number of common connections of i and j nodes with the neighboring ones and where we have assumed that the average connectivity is unaffected by the renormalization process (correct to leading order in $1/N$). Accordingly, $\Delta'-\Delta=\sigma-2-c$. If $c=0$ (no common connections), the average connectivity will increase (as σ is certainly larger than 2). In the opposite limit, if the two nodes have the same connections, $c=s_i-1$, $\Delta'-\Delta=\sigma-s_i-1$. In this case, the larger the s_i is, the larger the decrease is. From the numerical analyses, we can therefore infer that (above $T=0.02$) the nodes affected by the renormalization procedure are those characterized by a large connectivity.

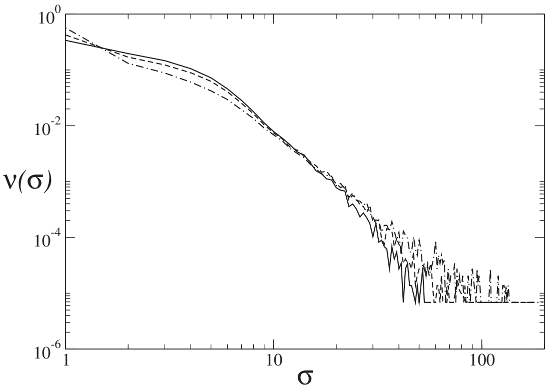

FIG. 3. Sequence S0: the fraction of nodes with connectivity σ, $\nu(\sigma)$, in log-log scale at $T=0$ (full line), $T=0.04$ (dashed line), and $T=0.08$ (dotted-dashed line).

In order to further clarify this issue, we have computed the distribution of the connectivity σ at different temperatures. In all cases, the distributions extend to around 100 connections (see Figs. 3–5). This limit is not simply due to the finite statistics but also to geometrical constraints: one indeed expects that the typical maximum number of neighbors is two times the dimension of the phase space (40 in our case). Larger values are possible but are increasingly unlike. For S0 (see Fig. 3) we also see that the distribution is approximately a power law and does not depend significantly on T (this is an obvious consequence of the very few renormalization steps that are involved). For S1 and S2, the zero temperature distribution is broader, but the nodes characterized by a high connectivity are progressively removed, thus confirming the previous interpretation of the dependence the average connectivity on T.

A special role is played by the "native" node, which progressively becomes the main hub of the network, as it turns out to include an increasing number of renormalized nodes as T increases. This effect is significantly more pronounced

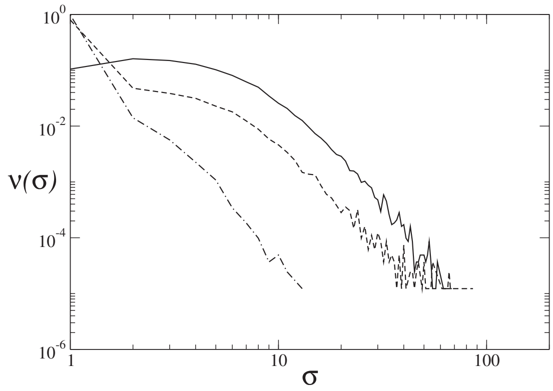

FIG. 4. Sequence S1: the fraction of nodes with connectivity σ, $\nu(\sigma)$, in log-log scale at $T=0$ (full line), $T=0.04$ (dashed line), and $T=0.08$ (dotted-dashed line).

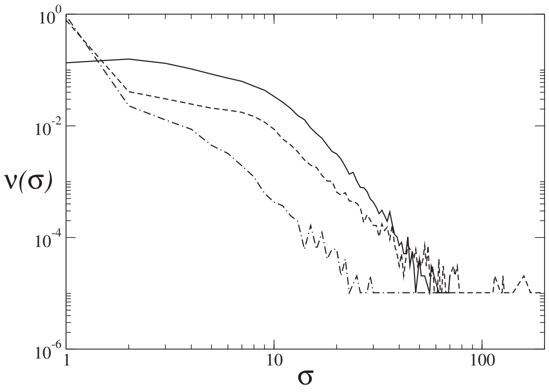

FIG. 5. Sequence S2: the fraction of nodes with connectivity σ, $\nu(\sigma)$, in log-log scale at T=0 (full line), T=0.04 (dashed line), and T=0.08 (dotted-dashed line).

in S1 and S2 (where its connectivity increases to $\simeq 10^4$) than in S0 (where it is "only" $\sim 10^3$).

Graphs can be also characterized in terms of the spectral dimension $\tilde{d}$ [25], which is defined from the formula

$$R(\omega) \sim \omega^{\tilde{d}} \quad \text{for} \quad \omega \to 0, \tag{22}$$

where $R(\omega)$ is the integrated density of eigenvalues of the topological Laplacian operator W_0 with eigenfrequency ω [29].

The dimension $\tilde{d}$ is well defined only in the limit of infinite graphs where ω can be arbitrarily small. However, also in cases like the present model, where the topological Laplacian matrix W_0 has a finite rank N (see Sec. II C), one can define an effective dimension. By denoting the eigenvalues of W_d with $\lambda_1 < \lambda_2 < \cdots < \lambda_k < \cdots < \lambda_N$, we can define the integrated density of eigenvalues $R(\lambda_k)$. By further identifying ω with $\omega_k = \lambda_k^{1/2}$, the spectral dimension can be estimated through the approximate relation,

$$R(\lambda_k) \sim \lambda_k^{\tilde{d}/2}, \tag{23}$$

which is again valid in the limit of small λ's. The identification of the full spectrum of the topological Laplacian W_0 requires to diagonalize a matrix of rank $O(10^5)$ (see Table II), a practically unfeasible task. Upon increasing the temperature, the rank of the renormalized discrete Laplacian matrix $\tilde{W}_0$ reduces (see Table VI), but its diagonalization remains a hard task for standard computational facilities. Fortunately, in order to estimate the spectral dimension it is sufficient to determine the lowest part of the Laplacian spectrum. This can be done by implementing Lanczos-like diagonalization algorithms such as those contained in the ARPACK library [30].

TABLE VI. Rank of the renormalized topological Laplacian matrix $\tilde{W}_0$ of the three investigated sequences at various temperatures below T_θ.

T	S0	S1	S2
0.02	138791	41139	38997
0.04	130326	20537	26292
0.06	117328	9430	15943
0.08	102933	4576	9736

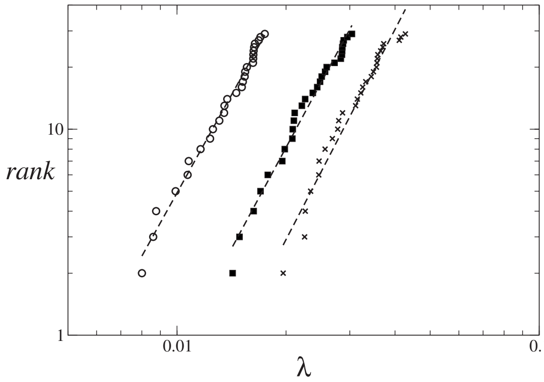

FIG. 6. Log-log plot of the spectrum of eigenvalues of the topological Laplacian matrix W_0 of the zero-temperature directed graph of S0 (empty circles), S1 (crosses), and S2 (filled squares). The dashed lines refer to power law with exponents of 3.3.

The data plotted in Fig. 6 show that the spectral dimension takes approximately the same value ($\tilde{d} \approx 6.5$) in all the three sequences considered in this paper. Moreover, in the two heteropolymers we see that the effective spectral dimension decreases to a value close to 5 for $T > 0.02$ (see Fig. 7, where we report the data for the sequence S1—a very similar scenario has been found for S2). This indicates that, above T=0.02, the renormalization procedure substantially modifies the structure of the directed graphs: regions of high connectivity collapse onto few nodes and the average connectivity is thereby reduced. This is consistent with what is shown in Fig. 2.

Altogether, the topological indicators analyzed in this section allow us to distinguish the homopolymer S0 from the heteropolymers. On the other hand, no significant difference between S1 and S2 has been detected. As discussed in Sec.

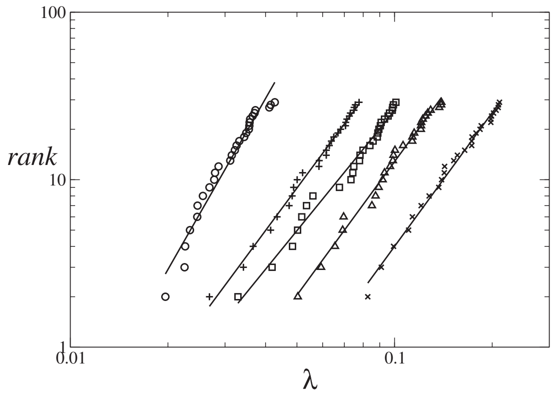

FIG. 7. Log-log plot of the spectrum of eigenvalues of the topological Laplacian matrix W_0 of S1 for five different temperatures: T=0.00 (○), 0.02(+), 0.04 (□), and 0.06 (△), 0.08(×). The data obtained for different temperatures are shifted horizontally by an arbitrary constant factor in order to obtain a better view. Notice that the power-law fit passes from its maximum value of 3.4 at T=0 to a minimum value of 2.5 above T=0.02.

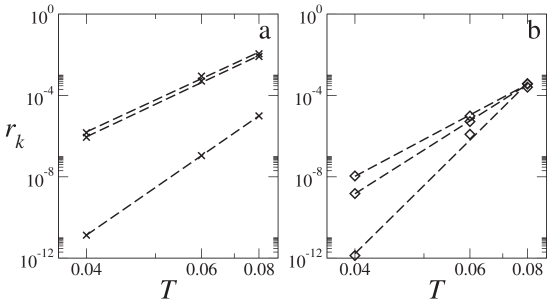

FIG. 8. The eigenvalues r_k, for $k=2,3,4$ versus temperature T in log-reciprocal scale for (a) S1 and (b) S2. The dashed lines are Arrhenius fits to the data.

III, this seems to be related to the dynamical properties of trajectories on the directed graphs. In Sec. IV B, we investigate the effect of the graph renormalization on the dynamics.

B. Relaxation times and the low-frequency spectrum of the Laplacian matrix

In this section we discuss the lowest part of the spectrum of Laplacian matrix (10) that characterizes the slowest relaxation processes. As discussed in Sec. IV A with reference to W_0, also in this case the accomplishment of this task requires using the Lanczos' diagonalization procedure.

The renormalization procedure described in Sec. IV A amounts to a series of transformations that involve the smallest barriers and therefore short time scales. Accordingly, the lowest part of the spectrum of W should be unaffected by the renormalization procedure. The numerical results for $T=0.08$ confirm this expectation (at lower temperatures, the diagonalization of W is too time or memory consuming, although $\widetilde{W}$ can be still analyzed).

In Fig. 8 we show the dependence of the smallest nonzero eigenvalues of $\widetilde{W}$ on the temperature T for the sequences S1 and S2. In both cases, there is a clear evidence of an Arrhenius-like behavior,

$$r_k \simeq A \exp(-B/k_B T), \tag{24}$$

where A and B are suitable constants, which depend on the sequence and on the eigenvalue. In particular B measures the effective height of the energy barrier.

The corresponding eigenvectors are shown in Figs. 9 and 10, where the components are ordered according to the energy of the corresponding graph node. The absolute values of the components of $w^{(2)}$, $w^{(3)}$, and $w^{(4)}$ are plotted together with $w^{(1)}$ that was reported for comparison. As shown in Appendix A, $w^{(i)}$, for $i \geq 2$, has zero average. With the only exception of $w^{(1)}$ (which describes the stationary distribution), all the eigenvectors of S2 are localized on some node of the graph, while only the second eigenvector $w^{(2)}$ of S1 is localized.

We have verified that for the localized eigenvectors the energy B corresponds to the height of the lowest energy barrier separating the localization node from the rest of the graph. This energy barrier is quite high so that the node can be viewed as a kinetic trap of GD. We have also found that most of the first 50 nonzero eigenvectors of S2 are localized as those shown in Fig. 9. Accordingly, many nodes do act as kinetic traps, thus slowing down many trajectories on the graph.

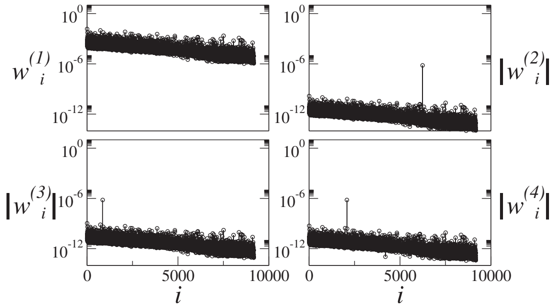

FIG. 9. Components of $w^{(1)}$ and absolute value of the components of $w^{(2)}$, $w^{(3)}$, and $w^{(4)}$ for the slow-folder S2 at $T=0.06$.

In the case of the delocalized eigenvectors of S1, B can be interpreted as an effective energy barrier separating different subsets of nodes in the graph. This interpretation is quite obvious for graphs composed of two weakly connected components. In this case, activation processes connect two set of nodes rather than two single minima. The involved nodes can be identified by dividing, e.g., each component $w_i^{(2)}$ by $w_i^{(1)}$ to obtain the "normalized" eigenvector $\overline{w}^{(2)}$. In fact, it turns out that the components of $\overline{w}^{(2)}$ are split into two sets of values, depending whether the corresponding node belong to the first or to second subgraph (see Appendix B for the details). The normalized eigenvectors $\overline{w^{(3)}}$ and $\overline{w^{(4)}}$, shown in the upper panels of Fig. 11, exhibit a similar behavior although they are split into more than two sets of values. This indicates that the graph structure of S1 is more intricated than in the example discussed in Appendix B. Nonetheless, the interpretation of B as an effective barrier height separating different regions of the graph remains valid. Notice also that the corresponding eigenvalues r_3 and r_4 are more than four orders of magnitude larger than r_2. This shows that the perturbative argument presented in the Appendixes A and B provides only a qualitative approximation of what was seen in Fig. 11.

For what concerns the folding process, we have already observed in Sec. III that the time scales of equilibration are orders of magnitude longer than those characterizing the first passage time t_f through the native valley (i.e., the native minimum and the connected minima). In fact, according to

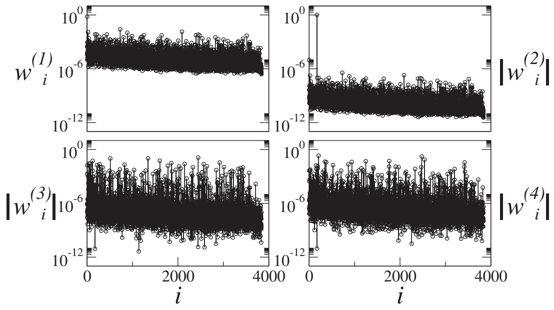

FIG. 10. Components of $w^{(1)}$ and absolute value of the components of $w^{(2)}$, $w^{(3)}$, and $w^{(4)}$ for the fast-folder S1 at $T=0.04$.

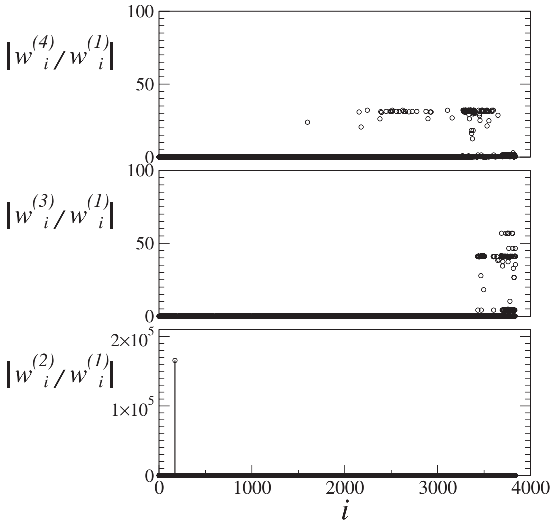

FIG. 11. Absolute value of the components of the normalized eigenvectors $\overline{w^{(2)}}$, $\overline{w^{(3)}}$, and $\overline{w^{(4)}}$ for the fast-folder S1 at $T=0.04$.

our definition of the folding temperature, we can guess that over the equilibration time scale approximately 50% of the time is spent in the native valley despite it contains a very small fraction of the minima (nodes) in the landscape (directed graph). In the renormalized representation of the directed graph, we expect that the quantitative determination of the average t_f through the native valley (that is reduced to a single node, as a result of the renormalization) is preserved, provided the graph is renormalized for temperatures $T \leq T_f$. For instance, we have verified that this is the case of S1 at $T=0.04$, where the folding time on the renormalized graph amounts to approximately 1900 units when averaged over 10^4 paths.

V. CONCLUSIONS AND PERSPECTIVES

Many phenomena of biological interest are associated with equilibrium and nonequilibrium processes in polypeptidic chains. A suitable description and understanding of such processes is far from trivial in these complex structures. This is the main reason why in this paper we have decided to consider a sufficiently simple and widely investigated model [16]. In particular, we have analyzed two heteropolymers and one homopolymer in order to clarify differences and analogies among various typical polypeptidic sequences. In fact, one of the heteropolymers is known to behave as a fast folder at variance with the other ones, which exhibit a much slower relaxation dynamics to their native states.

As a first step, we have described a strategy to reconstruct the PEL of these simple chains. The search for minima and first-order saddles has been performed by combining different algorithms aiming at a sufficiently careful reconstruction of the PEL close to the native minimum and up to energy values of the order of k_BT_f, where T_f denotes the folding temperature. Since the number of minima and saddles increases with the energy, the computational cost is already quite high up to k_BT_f: going beyond this value is practically impossible. On the other hand, performing a more extended search is not expected to add any relevant information. In fact, we have checked that for sufficiently low temperatures the main dynamical mechanisms associated with the folding process and with the relaxation to equilibrium are well reproduced even if most of the stationary points in the PEL above k_BT_f are discarded.

As a next step, we have constructed a directed graph representation of the dynamics, where the temperature-dependent molecular dynamics is replaced by a Markov-chain dynamics. The nodes of the graph correspond to the local minima of the PEL, while the first-order saddles connecting such minima are represented by the links of the graph. The strength of the links is measured by Langer's estimate [10,11] of the hopping rates between connected minima. We have shown that for temperatures close to or below T_f, MD simulations are essentially in good quantitative agreement with GD. However, in general, the latter approach systematically underestimates the hopping rates mainly because of the poorly reconstructed portion of the PEL above k_BT_f.

The main advantage of the graph representation is that one can apply a renormalization procedure that preserves the large-scale dynamical properties while cutting out many irrelevant degrees of freedom. In fact, the procedure described in Sec. IV allows to merge many nodes into single entities, most of which are characterized by a lower connectivity. The effect of the renormalization is more pronounced in heteropolymers than in the homopolymer, thus indicating that the topological properties are quite different in the two cases. Moreover, at least in the simple model considered in this paper, we have found evidence that topological indicators are not sufficient to discriminate between "fast" and "slow" folders. A clear distinction can instead be made by analyzing dynamical and spectral properties of the renormalized Laplacian matrix. Almost by definition, the first passage time t_f from the native configuration is expected to be much shorter in fast folders. In fact, we find that t_f is at least two orders of magnitude smaller in S1 than in S2. The interesting observation is that while in S2 t_f is comparable to the inverse of the smallest nonzero eigenvalue of the Laplacian matrix, in S1 t_f is three orders of magnitude shorter. This seemingly awkward result is due to the existence of a local minimum that is almost decoupled (i.e., it is separated by a high free-energy barrier) from the rest of the energy landscape. As, during the equilibration process, it is quite unlikely for the heteropolymer to be trapped in such a minimum, it is clear that it does not contribute significantly to slowing down of the folding process. On the other hand, its existence necessarily contributes to generating a very small low frequency in the Laplacian spectrum.

Moreover, we have found further important differences between the structures of the normalized eigenvectors $\bar{w}_i$ of S2 an S1. In the former case, all eigenvectors corresponding to the lowest eigenvalues are localized (at least, this is so for the first 100 ones). This indicates that the longer time scales are dominated by a "forest" of single-minimum kinetic traps in the energy landscape. Conversely, in S2, there is only one localized eigenvector that corresponds to a minimum where the dynamics could be accidentally trapped. All the other eigenvectors of S2 are extended, a fact which implies that they correspond to sets of multiple minima that have multiple mutual connections and are thus characterized by faster relaxation processes. It will be worth exploring to what extent this is a peculiarity of the simple model we have investigated in this paper or a more general signature of the differences between bad and good folders.

In fact, the methods described in this paper can be extended also to more realistic models of polypeptides and single-domain proteins. The reconstruction of a meaningful portion of the energy landscape will require higher computational costs and it will be necessary to devise specific techniques to identify the relevant minima. Nonetheless, there could be a positive payoff since the renormalization procedure would provide an effective characterization of the kinetics of these models, up to extremely long time scales, which would not be otherwise directly accessible. A first step in this direction has been attempted in [31].

ACKNOWLEDGMENTS

We acknowledge CINECA in Bologna and INFM for providing us access to the Beowulf Linux-Cluster under the grant "Iniziativa Calcolo Parallelo." This work was partially supported by the European Community via the STREP project EMBIO (NEST Contract No. 12835) and by the Italian project "Dinamiche cooperative in strutture quasi unidimensionali" No. 827 within the CNR program "Ricerca spontanea a tema libero."

APPENDIX A: PROPERTIES OF THE MASTER EQUATION

We will here review a few useful properties of the master equation [Eq. (8)] and those of eigenvectors of the Laplacian matrix that, although already widely reported in the literature, might help the reader in understanding the mathematical details of this paper.

As a preliminary observation we note that by summing both sides of the master equation over all the nodes of the graph and invoking the detailed balance condition, one can easily verify that the total probability is conserved,

$$\frac{d\left(\sum_i P_i\right)}{dt} = -\sum_{i,j=1}^{N} P_i \Gamma_{i\to j} + \sum_{i,j} P_j \Gamma_{j\to i} = 0. \qquad (A1)$$

The normalization condition $\Sigma_i P_i(t)=1$ therefore holds at every t, which, as we will see later, induces some constraints on the projections of realistic probability vectors on the eigenvectors of the Laplacian matrix W.

As already mentioned, W can be cast into a symmetric form through a similarity transformation. It therefore admits a complete basis of orthogonal eigenvectors, each describing a different mode of decay to equilibrium. Besides orthogonality, the eigenvectors of W share the additional property that their components are zero sum. In fact, from the eigenvalue equation $Ww_i^{(j)}=\lambda^{(j)}w_i^{(j)}$ one gets

$$w_i^{(j)} = \frac{-\sum_{k=1}^{N} w_i^{(j)}\Gamma_{i,k} + \sum_{k=1}^{N} w_k^{(j)}\Gamma_{k,i}}{\lambda^{(j)}}. \qquad (A2)$$

By summing both sides of this equation over i and using the detailed balance condition, one finds

$$\sum_{i=1}^{N} w_i^{(j)} = 0. \qquad (A3)$$

The only eigenvector that defies this demonstration is $w^{(1)}$, the null eigenvector defined in Eq. (11), which has positive components and can be normalized to unity. Using this normalization one can write

$$\sum_{i=1}^{N} w_i^{(j)} = \delta_{1,j}. \qquad (A4)$$

Actually, this condition is just a consequence of the fact that the master equation conserves probability. Indeed, since eigenvectors form a complete basis, each probability distribution of initial conditions on the graph $P(0)$ can be expressed as $P(0)=\Sigma_{j=1}^{N}\alpha_j w^{(j)}$. It will then evolve in time according to

$$P(t) = \sum_{j=1}^{N} \alpha_j w^{(j)} \exp^{-r_j t}. \qquad (A5)$$

Summing over the components of $P(t)$,

$$\sum_{i=1}^{N} P_i(t) = \sum_{j=1}^{N} \alpha_j \delta_{1,j} \exp^{-r_j t} = \alpha_0. \qquad (A6)$$

Hence, in order to have $\Sigma_{i=1}^{N} P_i(t)=1$, α_0 must necessarily be 1.

APPENDIX B: SPECTRAL CLUSTERING

We now justify the use of normalized components as an effective tool to uncover the inherent structure of eigenvectors. We define the normalized components of a vector v on a graph as the ratio site-to-site of the vector component to the local value of the stationary probability: $\bar{v}_i=v_i/w_i^{(0)}$. We will here extend an argument originally proposed for discrete graphs [32] to weighted ones and show that, when a graph is divided into two weakly connected subgraphs $\mathcal{A}$ and $\mathcal{B}$, the normalized components of the first nonzero eigenvector assume only two possible values, one for $\mathcal{A}$ and one for $\mathcal{B}$.

The Laplacian matrix W can be written as the sum of two matrices: $W=D-\Gamma^T$, where Γ is the transition rate matrix and D is a diagonal matrix, $D_{i,j}=\delta_{i,j}\Sigma_{k=1}^{N}\Gamma_{i,k}$. First of all, we consider the case in which the graph is composed of two disconnected subsets of nodes $\mathcal{A}$ and $\mathcal{B}$. In this case the Laplacian matrix will be referred to as W^0. Since $\Gamma_{i,j}=0$ for each $i\in\mathcal{A}$ and $j\in\mathcal{B}$, W^0 can be written as

$$W^0 = \begin{pmatrix} D_{\mathcal{A}\mathcal{A}} - \Gamma^T_{\mathcal{A}\mathcal{A}} & 0 \\ 0 & D_{\mathcal{B}\mathcal{B}} - \Gamma^T_{\mathcal{B}\mathcal{B}} \end{pmatrix}. \qquad (B1)$$

Let now use the null eigenvector $w^{(1)}$ of W^0 to construct two vectors $w^{\mathcal{A}}$ and $w^{\mathcal{B}}$ as follows:

$$w_i^{\mathcal{A}} = \begin{cases} w_i^{(1)}, & i \in \mathcal{A} \\ 0, & i \in \mathcal{B}, \end{cases} \qquad w_i^{\mathcal{B}} = \begin{cases} 0, & i \in \mathcal{A} \\ w_i^{(1)}, & i \in \mathcal{B}. \end{cases} \tag{B2}$$

It is easy to show that both $w^{\mathcal{A}}$ and $w^{\mathcal{B}}$ are eigenvectors of W^0 with a null eigenvector, and the same holds for any linear combination $v = aw^{\mathcal{A}} + bw^{\mathcal{B}}$. More generally when a graph is composed of n disconnected subgraphs the kernel of its Laplacian matrix has dimension n.

Let us now suppose that the two subgraphs $\mathcal{A}$ and $\mathcal{B}$ are not properly disconnect but do share a small number of connections. In this case the Laplacian matrix has the form $W = W^0 + W^1$, with

$$W^1 = \begin{pmatrix} D_{\mathcal{AB}} & -\Gamma_{\mathcal{AB}}^T \\ -\Gamma_{\mathcal{BA}}^T & D_{\mathcal{BA}} \end{pmatrix}, \tag{B3}$$

where $\Gamma_{\mathcal{AB}}$ and $\Gamma_{\mathcal{BA}}$ carry the information about connections between $\mathcal{A}$ and B, while $D_{\mathcal{AB}}$ and $D_{\mathcal{BA}}$ carry the information on the effect these connections have on the diagonal of the Laplacian matrix.

We now look for the eigenvectors of W among vectors of the form $v = aw^{\mathcal{A}} + bw^{\mathcal{B}}$,

$$Wv = (W^0 + W^1)(aw^{\mathcal{A}} + bw^{\mathcal{B}}) = W^1(aw^{\mathcal{A}} + bw^{\mathcal{B}}). \tag{B4}$$

In other words if any eigenvector of W exists, which is a linear combination of $w^{\mathcal{A}}$ and $w^{\mathcal{B}}$, it is also an eigenvector of W^1. We will therefore look for the eigenvectors of this last matrix having the desired form $aw^{\mathcal{A}} + bw^{\mathcal{B}}$.

For sufficiently small W^1 the vector $W^1(aw^{\mathcal{A}} + bw^{\mathcal{B}})$ can be approximately written as a linear combination of $w^{\mathcal{A}}$ and $w^{\mathcal{B}}$. To this purpose we introduce the projector on the space of the linear combinations of $w^{\mathcal{A}}$ and $w^{\mathcal{B}}$,

$$\Pi_{\mathcal{AB}} = (w^{\mathcal{A}}, w^{\mathcal{B}}), \tag{B5}$$

where $(w^{\mathcal{A}}, w^{\mathcal{B}})$ is a $N \times 2$ matrix whose columns are the two column vectors $w^{\mathcal{A}}$ and $w^{\mathcal{B}}$. The vector $W^1(aw^{\mathcal{A}} + bw^{\mathcal{B}})$ can now be written as

$$W^1(aw^{\mathcal{A}} + bw^{\mathcal{B}}) \simeq \Pi_{\mathcal{AB}} W^1 (aw^{\mathcal{A}} + bw^{\mathcal{B}}) = \widetilde{W}^1 \begin{pmatrix} a \\ b \end{pmatrix} (w^{\mathcal{A}}, w^{\mathcal{B}}), \tag{B6}$$

where we have introduced $\widetilde{W}^1 = \Pi_{\mathcal{AB}}^T W^1 \Pi_{\mathcal{AB}}$, a 2×2 matrix that reproduces the effect of W^1 in the subspace of the linear combinations of $w^{\mathcal{A}}$ and $w^{\mathcal{B}}$. After some algebra $\widetilde{W}^1$ reads as follows:

$$\widetilde{W}^1 = \begin{pmatrix} \sum\limits_{i\in\mathcal{A},\ j\in\mathcal{B}} w_i^{(1)^2}\Gamma_{i,j} & -\sum\limits_{i\in\mathcal{A}, j\in\mathcal{B}} w_i^{(1)} w_j^{(1)} \Gamma_{j,i} \\ +\sum\limits_{i\in\mathcal{B},\ j\in\mathcal{A}} w_i^{(1)} w_j^{(1)} \Gamma_{j,i} & \sum\limits_{i\in\mathcal{B},\ j\in\mathcal{A}} w_i^{(1)^2}\Gamma_{i,j} \end{pmatrix}. \tag{B7}$$

Since $w^{(1)}$ satisfies the detailed balance, and existing a connection from $\mathcal{B}$ to $\mathcal{A}$ for each connection from $\mathcal{A}$ to $\mathcal{B}$, one has

$$\sum_{j\in\mathcal{B}} w_i^{(1)}\Gamma_{i,j} = \sum_{j\in\mathcal{B}} w_j^{(1)}\Gamma_{j,i}, \quad \forall\, i \in \mathcal{A}. \tag{B8}$$

Analogously,

$$\sum_{j\in\mathcal{A}} w_i^{(1)}\Gamma_{i,j} = \sum_{j\in\mathcal{A}} w_j^{(1)}\Gamma_{j,i}, \quad \forall\, i \in \mathcal{B}. \tag{B9}$$

We can therefore define two quantities $\alpha = \Sigma_{i\in\mathcal{A},\ j\in\mathcal{B}} w_i^{(1)} w_j^{(1)} \Gamma_{j,i}$ and $\beta = \Sigma_{i\in\mathcal{B},\ j\in\mathcal{A}} w_i^{(1)} w_j^{(1)} \Gamma_{j,i}$ that cast $\widetilde{W}^1$ in a particularly simple form

$$\widetilde{W}^1 = \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}. \tag{B10}$$

It is important to notice that if the two subgraphs $\mathcal{A}$ and $\mathcal{B}$ are weekly connected α and β will be relatively small since there are few connections such that $\Gamma_{j,i} > 0$ for $i \in \mathcal{A}$ and $j \in \mathcal{B}$.

The two eigenvalues of $\widetilde{W}^1$ are 0 and $\alpha + \beta$, referring to the eigenvectors

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \alpha \\ -\beta \end{pmatrix}, \tag{B11}$$

respectively. It can finally be shown that by backprojecting with $\Pi_{\mathcal{AB}}^T$ these two eigenvectors of $\widetilde{W}^1$ to the entire N-dimensional space one obtain two eigenvectors of W^1 characterized by the same eigenvalues. According to Eq. (B4) these are also eigenvectors of W. More precisely by this procedure one obtains the following:

(i) $w^{\mathcal{A}} + w^{\mathcal{B}}$ that obviously coincides with $w^{(1)}$ and is associated to the null eigenvalue also according to the perturbative calculation and

(ii) $u = \alpha w^{\mathcal{A}} - \beta w^{\mathcal{B}}$, associated to the eigenvalue $\alpha + \beta$ that is a small number and will therefore lay in the end of the spectrum of W.

It is now straightforward to verify that the normalized coordinates of u get the values α for nodes belonging to $\mathcal{A}$ and $-\beta$ for nodes belonging to $\mathcal{B}$. In this sense the analysis of the normalized coordinates of the eigenvectors of W can be employed as a spectral method for the identification of clusters, portion of the graph characterized by a high degree of internal connectivity, and a small number of connections with the rest of the graph.

[1] R. Schilling, in *Collective Dynamics of Nonlinear and Disordered Systems*, edited by G. Radons, W. Just, and P. Häussler (Springer, Berlin, 2005).

[2] M. Dobson, A. Šali, and M. Karplus, Angew. Chem., Int. Ed. **37**, 868 (1998).

[3] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II: Nonequilibrium Statistical Mechanics* (Springer, Berlin, 2008).

[4] U. Balucani and M. Zoppi, *Dynamics of the Liquid State* (Clarendon, Oxford, 1994).

[5] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).

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

[7] F. H. Stillinger and T. A. Weber, Science **225**, 983 (1984).

[8] S. Sastry, P. G. Debenedetti, and F. H. Stillinger, Nature (London) **393**, 554 (1998).

[9] P. N. Mortenson and D. J. Wales, J. Chem. Phys. **114**, 6443 (2001).

[10] J. S. Langer, Ann. Phys. **54**, 258 (1969).

[11] P. Hänggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251 (1990).

[12] L. Bongini, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **68**, 061111 (2003).

[13] L. Bongini, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **72**, 051929 (2005).

[14] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).

[15] A. Torcini, R. Livi, and A. Politi, J. Biol. Phys. **27**, 181 (2001).

[16] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993).

[17] A. Irbäck, C. Peterson, and F. Potthast, Phys. Rev. E **55**, 860 (1997).

[18] Y. Jabri, *The Mountain Pass Theorem: Encyclopedia of Mathematics and its Applications* (Cambridge University Press, Cambridge, 2003).

[19] D. Moroni, T. S. van Erp, and P. G. Bolhuis, Physica A **340**, 395 (2004).

[20] J. P. K. Doye and D. J. Wales, Z. Phys. D: At., Mol. Clusters **40**, 194 (1997).

[21] L. J. Lewis and N. Mousseau, Comput. Mater. Sci. **20**, 285 (2001).

[22] The procedure can be accelerated by applying quasi-Newtonian algorithms [33].

[23] S. Gerschgorin, Izv. Akad. Nauk USSR Otd. Fiz.-, Mat. Nauk **7**, 749 (1931).

[24] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).

[25] R. Burioni and D. Cassi, Phys. Rev. Lett. **76**, 1091 (1996).

[26] M. A. Miller and D. J. Wales, J. Chem. Phys. **111**, 6610 (1999).

[27] S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002); D. A. Evans and D. J. Wales, *ibid.* **118**, 3891 (2003).

[28] F. Rao and A. Calfisch, J. Mol. Biol. **342**, 299 (2004).

[29] Notice that the low-frequency component of the spectrum corresponds to collective modes as in the case of network elements connected by harmonic interactions. In the case of a regular network, $\tilde{d}$ coincides with the lattice dimension d.

[30] D. C. Sorensen, Institute for Computer Applications in Science and Engineering (ICASE) Technical Report TR-96-40, 1996 (unpublished).

[31] M. Baiesi, L. Bongini, L. Casetti, and L. Tattini, Phys. Rev. E (to be published).

[32] C. H. Q. Ding, X. He, and H. Zha, Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, 2001 (unpublished), p. 275.

[33] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 2nd ed. (Wiley, New York, 2001).