Exploring the energy landscape of model proteins: A metric criterion for the determination of dynamical connectivity

Lorenzo Bongini,[1,2,*] Roberto Livi,[1,2,†] Antonio Politi,[2,3,‡] and Alessandro Torcini[2,3,§]

[1]*Dipartimento di Fisica, Universitá di Firenze, via Sansone, 1-I-50019 Sesto Fiorentino, Italy*
[2]*Centro Interdipartimentale per lo Studio delle Dinamiche Complesse, via Sansone, 1-I-50019 Sesto Fiorentino, Italy*
[3]*Istituto dei Sistemi Complessi, Consiglio Nazionale delle Ricerche, L.go E. Fermi 6 I-50125 Firenze, Italy*

A method to reconstruct the energy landscape of small peptides is presented with reference to a two-dimensional off-lattice model. The starting point is a statistical analysis of the configurational distances between generic minima and directly connected pairs (DCP). As the mutual distance of DCP is typically much smaller than that of generic pairs, a metric criterion can be established to identify the great majority of DCP. Advantages and limits of this approach are thoroughly analyzed for three different heteropolymeric chains. A funnel-like structure of the energy landscape is found in all of the three cases, but the escape rates clearly reveal that the native configuration is more easily accessible (and is significantly more stable) for the sequence that is expected to behave as a real protein.

I. INTRODUCTION

Several states of matter are characterized by a rich energy landscape, which, in turn, hints at peculiar structural and dynamical features. Supercooled liquids, glasses, atomic clusters, and biomolecules [1] are typical examples of systems whose complex thermodynamic behavior can be traced back to the intricate topological properties of the energy landscape. The pioneering work by Stillinger and Weber on "inherent" structures of liquids [2] revealed the importance of investigating the stationary points of the potential energy for characterizing their dynamical and thermodynamic properties. Similar approaches have been proposed and successfully applied to the identification of the structural-arrest temperature in glasses [3] and supercooled liquids [4].

More recently, this kind of analysis has been extended to the study of protein models [5–7]. They suggest that also the folding process of a protein toward its native configuration depends on the structure of its energy landscape. This has been found to possess a funnel-like shape: the native configuration is located inside the so-called native valley at the bottom of the funnel [8].

Below the folding temperature, the evolution from a coil state to the native configuration is determined by the propensity of the protein to enter the relatively small fraction of states composing the native valley without having to visit the entire phase space. In a statistical sense, the folding process can be viewed as a weighted sampling mechanism which favors specific intermediate configurations. They correspond to assembling the building blocks which eventually constitute the native topology. Well above the folding temperature, no marked difference exists among the various states and the protein spends most of the time jumping between different random coil configurations.

In the absence of external forces, only thermal fluctuations can drive the protein dynamics through different regions of the energy landscape. In particular, below the folding temperature, the protein is expected to evolve mainly inside the native valley. Nonetheless, large deviations from the native configuration cannot be avoided, but they are both rare and very short lived. This scenario was confirmed by simulations performed in a two-dimensional (2D) off-lattice model [9]. A more detailed analysis of the model [10] revealed that the protein dynamics can viewed as a sequence of jumps between pairs of minima separated by one saddle, that we call directly connected pairs (DCP). Each jump is a thermally activated process: the protein performs random oscillations in the basin of a local minimum until a sufficiently large thermal fluctuation allows it to overtake the energy barrier separating the minimum from a neighboring one. In a high-dimensional space, the transition rate can be computed as the product of the Arrhenius factor times an entropic weight which depends on the DCP and on the saddle curvatures (see Sec. IV). In other words, the relevant information about the protein dynamics can be obtained from the knowledge of the DCP and of the corresponding saddles. However, the reconstruction of the energy landscape is a very difficult task to be accomplished in practice. Indeed, the identification of DCP by a systematic exploration of the entire set of the N identified minima requires exploring N^2 pairs, which is already on the order of 10^9 for the partial database generated (see the Appendix for a description of the algorithm) in the relatively small polypeptidic chains investigated in this paper (see Sec. II). It is, therefore, crucial to develop effective strategies for identifying DCP within the set of all, *a priori* possible, candidates. This is the main issue addressed in this paper.

It is reasonable to conjecture that the distance separating DCP is typically smaller than that between generic pairs of minima. It is, therefore, tempting to restrict the analysis to

---
*Electronic address: bongini@fi.infn.it
†Electronic address: livi@fi.infn.it
‡Electronic address: politi@inoa.it
§Electronic address: torcini@inoa.it

those pairs whose mutual distance is smaller than some pre-scribed threshold. However, whether this approach effectively works may depend on several factors, one of which is the adopted definition of distance. For this reason, in Sec. III, we introduce and compare different conformational distances. It turns out that the bond-angle distance, defined by the absolute-value norm is the one that makes the implementation of such a criterion most accurate in the identification of DCP. The price one has to pay is that, unavoidably, all DCP characterized by a distance larger than the given threshold are missed. The detailed analysis of the energy landscape described in Sec. IV indicates that the average distance between DCP is larger for minima that are closer to the native topology. Accordingly, the computational advantage guaranteed by the choice of a relatively small threshold could be frustrated by the loss of important connections located in the native valley. For this reason, we argue that the distance criterion has to be complemented by a systematic search of all DCP involving minima in the native valley, a much more accessible task, given the limited number of such minima (for an operative definition of the native valley, see Sec. IV).

II. THE MODEL

The model studied in this paper is a modified version of the 2D off-lattice model introduced by Stillinger *et al.* [11] and already investigated in Ref. [9]. It consists of a chain of L point-like monomers mimicking the residues of a polypeptidic chain. For the sake of simplicity, only two types of residues are considered: hydrophobic (H) and polar (P) ones. Any chain is unambiguously identified by a sequence of binary variables $\{\xi_i\}$ ($i=1,\ldots,L$), where $\xi_i = \pm 1$ corresponds to H and P residues, respectively. The intramolecular potential is composed of three terms: a stiff nearest-neighbor harmonic potential V_1 intended to maintain the bond distance almost constant, a three-body interaction V_2 which accounts for the bending energy, and a long-range Lennard-Jones interaction V_3 acting on all pairs i, j such that $|i-j|>1$

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

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

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

Here, $r_{i,j}$ is the distance between the ith and the jth monomer and θ_i is the bond angle at the ith monomer. The parameters $\alpha=20$ and $r_0=1$ (both expressed in adimensional arbitrary units) fix the strength of the harmonic force and the equilibrium distance between subsequent monomers (which, in real proteins, is on the order of a few angstroms). The value of α is chosen to ensure a value for V_1 much larger than the other terms of potential (1) in order to reproduce the stiffness of the protein backbone. V_3 is the only term of the potential energy which depends on the nature of the monomers: the parameters $c_{i,j} = \frac{1}{8}(1+\xi_i+\xi_j+5\xi_i\xi_j)$ are chosen in such a way that the interaction is attractive if both residues are either

hydrophobic or polar ($c_{i,j}=1$ and $1/2$, respectively), while it is repulsive if the residues belong to different species ($c_{ij}=-1/2$).

Accordingly, the Hamiltonian of the system reads

$$H = \sum_{i=1}^{L} \frac{p_{x,i}^2 + p_{y,i}^2}{2} + \sum_{i=1}^{L-1} V_1(r_{i,i+1}) + \sum_{i=2}^{L-1} V_2(\theta_i)$$
$$+ \sum_{i=1}^{L-2} \sum_{j=i+2}^{L} V_3(r_{ij}, \xi_i, \xi_j), \qquad (2)$$

where, for the sake of simplicity, all monomers are assumed to have the same unitary mass and the momenta are defined as $(p_{x,i}, p_{y,i}) \equiv (\dot{x}_i, \dot{y}_i)$. Despite its simplicity, this toy model of a heteropolymer does reproduce the expected properties of polypeptidic chains and is thus very useful for testing dynamical and statistical indicators. For instance, accurate Monte-Carlo simulations, performed by employing innovative schemes [12], have revealed that, in analogy with real proteins, only a few sequences systematically fold onto the same native structure: this is why they have been named "good folders" [13,14]. Such studies have been confirmed and complemented by direct molecular dynamics simulations [9].

In this paper, we limit ourselves to investigating the three following sequences of 20 monomers, (i) [S0] a homopolymer composed of 20 H residues; (ii) [S1]=[HHHP HHHP HHHP PHHP PHHH] a sequence that has been identified as a good folder in [14]; (iii) [S2]=[PPPH HPHH HHHH HHHP HHPH] a randomly generated sequence that has been identified as a bad folder in [9].

These sequences have been chosen because they represent the three classes of different folding behaviors observed in this model. The main thermodynamic features of all of them can be summarized with reference to three different transition temperatures [10]. Decreasing the temperature from high values, one first encounters the temperature T_θ below which the sequence is typically found in a collapsed configuration, rather than in a random-coil one [15]. Then, one finds the so-called folding temperature T_f below which the heteropolymer stays predominantly in the native valley. Finally, at even lower temperatures, one finds T_g: this is the glass-transition temperature, below which a structural arrest of the system occurs.

In the following, we aim at a deeper understanding of the folding process by investigating directly the structure of the energy landscape of all these sequences.

III. ANALYSIS OF THE ENERGY LANDSCAPE

As pointed out in the introduction, the main problem for reconstructing the energy landscape amounts to finding DCP and the corresponding saddles of the potential energy $V=V_1+V_2+V_3$ [see Eq. (2)]. An exhaustive search of DCP among all pairs of minima rapidly becomes unfeasible with increasing the chain length L, due to the exponential increase of the number of minima with L itself [1]. Since the total number of DCP is a rather small fraction of all possible pairs

TABLE I. Number of minima in the whole database and in the first and second shell. Number of investigated pairs and number of DCP found. Data have been reported for all the three examined sequences.

	S0	S1	S2
Total number of minima	2.3×10^5	7.2×10^4	6.4×10^4
Number of minima in the first shell	64	50	49
Number of minima in the second shell	1181	465	348
Number of minima below T_f	66 470	5883	8670
Number of investigated pairs	9.1×10^6	0.94×10^6	1.3×10^6
Number of connected pairs	84 990	10 470	14 356

(see, e.g., Table I), it would be very helpful finding an effective criterion to restrict the search of potentially directly connected minima. *A priori*, the *distance* seems to be the right indicator to discriminate between connected and not-contiguous pairs of minima. In this section, we investigate several definitions of distance with the goal of identifying the most appropriate one to identify DCP.

As a first candidate, we introduce the generalized angular distance $\delta_\theta^{(q)}(\mathcal{C}_1, \mathcal{C}_2)$ between configurations $\mathcal{C}_1$ and $\mathcal{C}_2$,

$$\delta_\theta^{(q)}(\mathcal{C}_1, \mathcal{C}_2) = \left(\frac{1}{L-2} \sum_{i=2}^{L-1} |\theta(i;\mathcal{C}_1) - \theta(i;\mathcal{C}_2)|^q \right)^{1/q}, \quad (3)$$

where $\theta(i, \mathcal{C})$ is the ith bond angle of the configuration $\mathcal{C}$. Notice that this angular distance is very sensitive to local fluctuations along the chain. A generalized distance which depends more on the global than on the local structure of a configuration is

$$\delta_r^{(q)}(\mathcal{C}_1, \mathcal{C}_2) = \left(\frac{2}{L(L-1)} \sum_{i>j+1}^{L} |r(i,j;\mathcal{C}_1) - r(i,j;\mathcal{C}_2)|^q \right)^{1/q}, \quad (4)$$

where $r(i,j;\mathcal{C})$ is the intrabead distance between the ith and jth monomers of the configuration $\mathcal{C}$. This distance is related to the χ indicator, previously employed in the analysis of the folding dynamics in on-lattice and off-lattice models of heteropolymers in 2D and three dimensions (3D) [9,16]. For $q=1$, 2, and $+\infty$, both definitions of generalized distances reduce to the standard absolute-value, Euclidean, and maximum norms, respectively. Such distances have been computed for all the pairs of minima in the databases of the sequences S0, S1, and S2. The algorithm used to generate the databases is described in the Appendix. We want to point out that any numerical procedure, including ours, cannot guarantee the identification of all minima and saddles in the energy landscape. Nonetheless, we have independently verified that
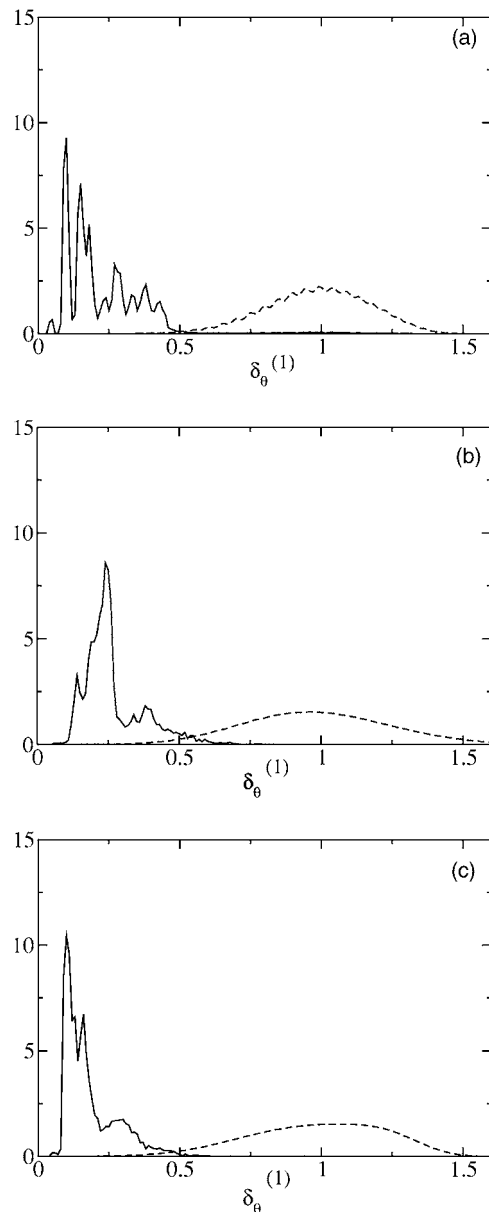

FIG. 1. Probability density of angular distances $\delta_\Theta^{(1)}$ for all the pairs of minima, P_a (dashed line), and for DCP, P_c (solid line), for the sequences S0 (a) and S1 (b) and S2 (c).

the algorithm allows obtaining at least a very accurate description of the native valley.

Then, we have computed the probability densities of the generalized distances between generic $[P(\delta)]$ and directly connected $[P_c(\delta)]$ pairs of minima. In all cases, P has a bell shape with a maximum close to 1, while P_c is sharply peaked at much smaller values (see, e.g., Fig. 1, where $P(\delta_\theta^{(1)})$ and $P_c(\delta_\theta^{(1)})$ are plotted for the sequences S0, S1, and S2). This confirms the naive idea that DCP are typically much closer than randomly chosen pairs of minima.

A qualitatively similar difference between P and P_c is observed also for different choices of q as well as for the global distance $\delta_r^{(q)}$. In order to identify the most appropriate value of q, it is convenient to introduce the integrated fraction $R(\delta)$ of pairs of minima whose distance is smaller than δ
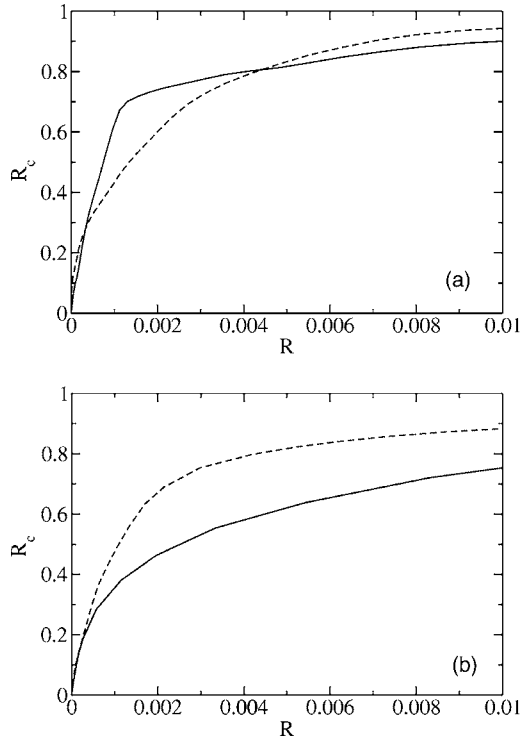
FIG. 2. Integrated probability density R_c versus R for the angular δ_Θ (solid lines) and global δ_r (dashed lines) distances. The data refer to the sequence S1 and to the absolute value (a) and maximum (b) norms.

$$R(\delta) = \int_0^\delta P(x)\mathrm{d}x \qquad (5)$$

and, equivalently,

$$R_c(\delta) = \int_0^\delta P_c(x)\mathrm{d}x, \qquad (6)$$

relative to DCP. Upon considering δ as a dummy variable, it is possible to plot R_c versus R. A fast convergence of R_c to 1 means that almost all DCP can be already identified by limiting the search to relatively close pairs of minima. The data reported in Fig. 2 for the sequence S1 indeed reveal that such a fast convergence of R_c is achieved for both angular and global distances. However, regardless of the chosen distance, the absolute value norm ($q=1$) represents a more effective indicator of connectivity between DCP than the maximum norm ($q=\infty$). On the other hand, Fig. 2(a) shows that for large R, $\delta_r^{(1)}$ performs slightly better than $\delta_\theta^{(1)}$, which conversely is definitely more effective for small R. This is why in what follows we limit our analysis to $\delta_\theta^{(1)}$.

In order to clarify the role of the parameter q, in Figs. 3 and 4, we have plotted R_c versus R for different values of this parameter. There we see that, upon decreasing q from ∞ down to 1, R_c exhibits an increasingly fast saturation, while the opposite is observed when q is further decreased below 1. The bad performance observed at high q values has a quite intuitive explanation: in that limit, the norm reduces to the maximum norm and the slow growth of R_c tells us that dis-
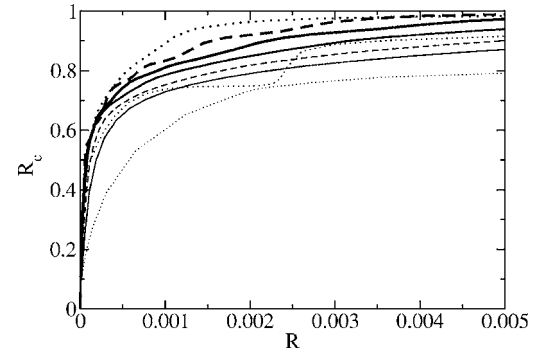

FIG. 3. Integrated probability density R_c versus R for the angular distance for various norms for the sequence S0. Continuous lines correspond to $q < 1$ ($q=0.01$, $q=0.2$, $q=0.5$, and $q=0.75$, going from the thinnest to the tickest line, respectively). The dashed line corresponds to $q=1$ (norm of the absolute value). The dotted lines correspond $q > 1$ ($q=2$, $q=5$, and $q=\infty$, going from the thickest to the thinnest line, respectively).

tances between DCP are not uniformly small along all directions: DCP may significantly differ along specific directions in spite of being "on the average" much closer than generic pairs of minima. The relatively bad performance observed for $q \rightarrow 0$ has a complementary explanation. In that limit, the average distance is strongly biased by small differences $\delta\theta$ or δr, whose occasional occurrence may induce us to classify as "close," configurations that are significantly different instead. In all configurations, we have investigated, it turns out that $q \approx 1$ is the best compromise between the above two effects. Having established that $q=1$ is the best choice, from now on,
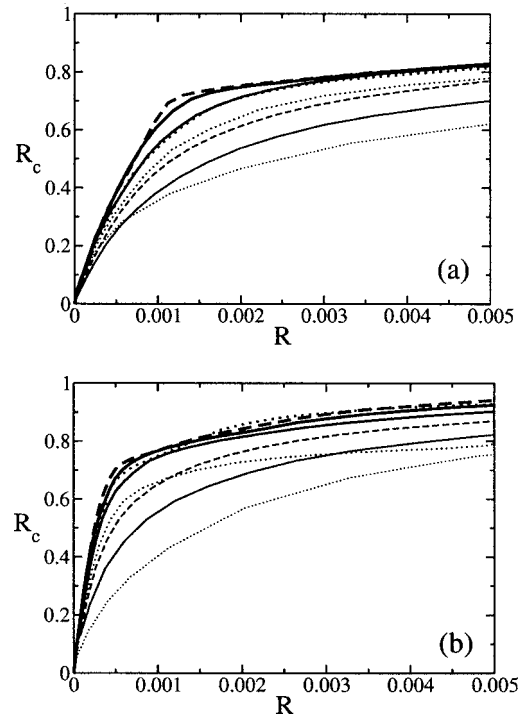

FIG. 4. Integrated probability density R_c versus R for the angular distance for various norms for the sequences S1 (a) and S2 (b). The notations are the same as in Fig. 3.
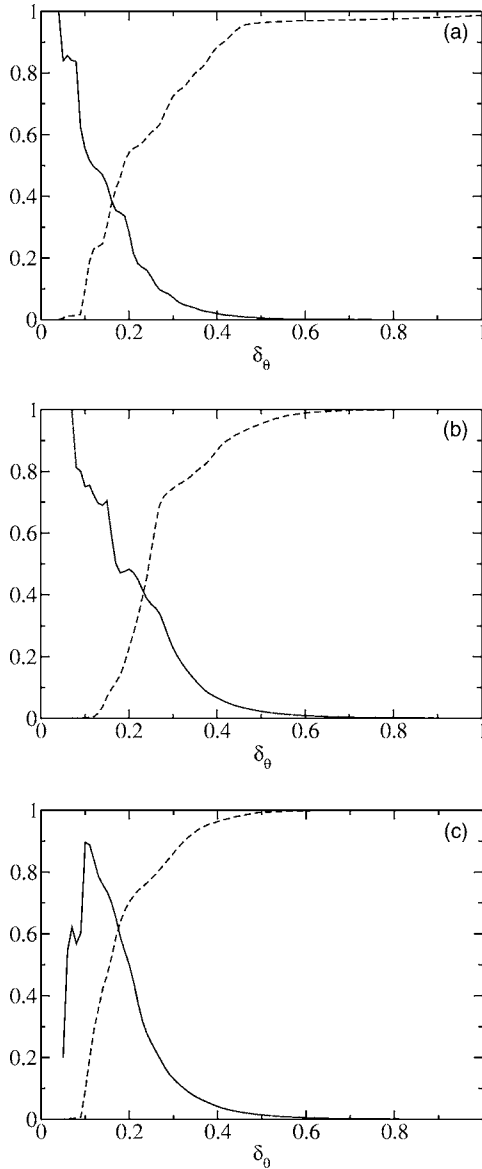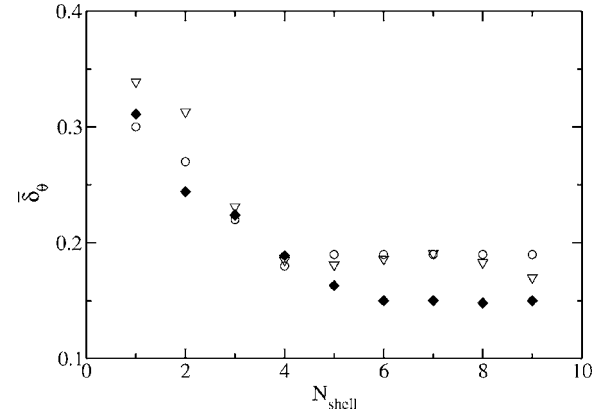
FIG. 6. Average connection length between the minima of the ith and $(i-1)$th shell for the three sequences S0 (triangles), S1 (circles), and S2 (filled diamonds).

most time-consuming part of the systematic search algorithm, which amounts to testing whether a generic pair of minima is separated by a single saddle.

The drawback is that for any choice of $\delta_\theta^\star$, those DCP whose mutual distance is anomalously large are going to be missed. In Sec. IV, we analyze the energy landscape with the main goal of concluding whether such a component is of some relevance in the overall reconstruction of the energy landscape.

IV. A CLOSER INSPECTION OF THE ENERGY LANDSCAPE

A faithful reconstruction of the energy landscape requires a sufficiently large database of stationary points, i.e., minima and saddles. The procedure described in the Appendix is quite reliable in this respect, but it is also computationally very time-consuming as already stated. In Sec. III, we have seen that a large fraction of DCP can be obtained by adopting a suitable metric criterion. However, it is not *a priori* clear whether the long-distance tail of P_c is qualitatively irrelevant too.

In order to shed some light on this question, we have divided the minima into "shells": the nth shell is defined as the collection of all minima which are separated from the native configuration by at least n saddles. The identification of the minima belonging to each shell can be achieved recursively: (i) the 0th shell coincides with native configuration; (ii) a minimum $\mathcal{C}_1$, directly connected to a minimimum $\mathcal{C}_2$ of the ith shell, is identified as part of the $(i+1)$st shell if $\mathcal{C}_2$ does not belong to a shell of order $j \leqslant i$.

In Fig. 6, we report the average value $\bar{\delta}_\theta$ of the distance between DCP belonging to consecutive shells for all of the three sequences. This figure indicates that the interminimum distance grows in the vicinity of the native configuration. The average distance $\langle \delta_\theta \rangle$ between DCP lying inside the same shell exhibits the same behavior. Finally, this rarefied density of minima in the vicinity of the native configuration is confirmed also by plotting the mutual distance versus the actual distance from the native configuration (native dis-

FIG. 5. The ratio ρ between DCP and all pairs up to distance δ_Θ (continuous line) compared with $R_c(\delta_\Theta)$ (dashed line) for the three analyzed sequences, S0 (a), S1 (b), and S2 (c).

we limit ourselves to considering that value and drop the superscript (1) in the definition of the distance.

In practice, since it is eventually necessary to identify a threshold distance $\delta_\theta^\star$, it is convenient to look also at the dependence of R_c on δ_θ and to introduce

$$\rho(\delta_\theta) = \frac{R_c(\delta_\theta)}{R(\delta_\theta)}.$$

From Fig. 5, we see that $\delta_\Theta^\star = 0.5$ is a good choice, since $\rho(0.5) \sim O(10^{-2})$, while $R_c(0.5) \sim 99\%$. Even reducing the threshold value to $\delta_\theta^\star = 0.2$, a large fraction of DCP are still recovered $[R_c(0.2) \sim 90\%]$.

These results indicate that if one restricts the search of DCP to the set of minima whose distance is smaller than a prescribed threshold $\delta_\theta^\star$, one can reduce significantly the
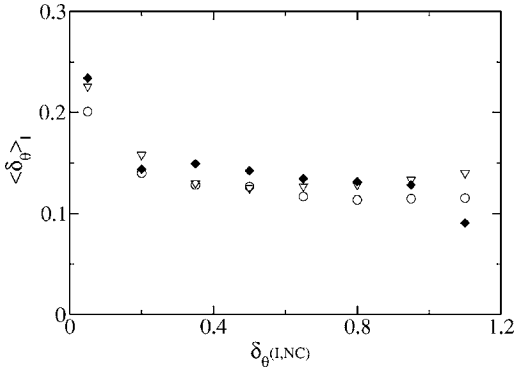
FIG. 7. The average distance of minima from their connected set versus their native distance for the sequence S0 (triangles), S1 (circles), and S2 (filled diamonds). The data are averaged also over intervals of length $I = 0.01$ along the horizontal axis.

tance). In this case, in order to smoothen the wild pair-to-pair fluctuations, a coarse graining has been performed by averaging over bins of width 0.01 along the δ_θ axis (see Fig. 7).

Altogether, the significant increase of the average distance close to the native valley indicates that some of the DCP that are missed by the metric criterion discussed in Sec. III lie in the most relevant region for the characterization of the folding and/or unfolding processes. However, considering the limited number of minima in the native valley, such a difficulty can be easily overcome by complementing the overall application of the metric criterion with an extensive comparison of such minima with all configurations in the database: the additional cost in terms of the computing time is indeed a minor one.

The increased distance between neighboring minima hints at possibly deeper valleys and is, in turn, suggestive of the presence of a funnel in the energy landscape, a structure that is typically expected to appear in true proteins [8]. However, we find this scenario in all of the three sequences analyzed in this paper, including the homopolymer S0 which cannot be certainly considered a reasonable model for a protein. In order to further clarify this point, we have computed the escape rates from the single valleys. Given any two directly connected minima $\mathcal{C}_1$ and $\mathcal{C}_2$ characterized by the potential energies V_1 and V_2 ($V_1 \leqslant V_2$), the escape rate from the minimum $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2\}$ through the saddle $\mathcal{C}_s$ (with potential energy V_s) is given by

$$\Gamma_{\mathcal{C}} = \frac{\omega_\parallel}{\pi \gamma} \frac{\prod_{i=1}^{L'} \omega_{\mathcal{C}}^{(i)}}{\prod_{i=1}^{L'-1} \omega_\perp^{(i)}} \exp\left\{ -\frac{W_{\mathcal{C}}}{k_B T} \right\}. \tag{7}$$

This formula, proposed by Langer [17,18], is obtained by considering the harmonic approximation of the potential in the vicinity of $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_s$. The $\omega_{\mathcal{C}}^{(i)}$'s are the $L' = 2L - 3$ nonzero frequencies of the minimum $\mathcal{C}$ ($\omega_{\mathcal{C}}^{(i)} = \sqrt{-\Lambda_{\mathcal{C}}^{(i)}}$, where $\Lambda_{\mathcal{C}}^{(i)}$ is the ith negative eigenvalue of the Hessian of the potential energy V). Analogously, $\omega_\perp^{(i)}$'s are the $L' - 1$ nonzero frequencies of the saddle (those corresponding to the stable directions), while $\omega_\parallel$ is associated with the only expanding direction. Finally, γ is the dissipation rate [19], while the

Arrhenius exponential factor depends on the height of the barrier, $W_{\mathcal{C}} = V_s - V_{1,2}$, normalized to the reduced temperature $K_B T$, K_B being the Boltzmann constant.

The above expression has been shown to reproduce reasonably well the numerically obtained escape rates for heteropolymers in 2D at moderate temperatures for $T \lesssim T_\theta$ [10]. Accordingly, in that regime, both the folding and the unfolding dynamics toward and from the native configuration are driven by thermal activation processes, which determine the transitions between DCP. Small Γ values suggest that the heteropolymer may be trapped into some local valley far from the native configuration. Equation (7) shows explicitly that in high-dimensional spaces, the escape rate does not simply depend on the energy barriers but also on entropic factors, which depend on geometrical features of basins of attractions of the stationary configurations in the energy landscape.

In Fig. 8, we have plotted the rates $\Gamma_{\mathcal{C}}$ as a function of δ_θ for the three sequences at $T = T_f$ and $\gamma = 1$ (since past simulations indicate that the effective value of γ is the same at least in the whole native valley, there is no need to know it when a comparative analysis is being carried out). There we notice a striking difference between S0 and S1 at large distances: actually the escape rates of S0 are two orders of magnitude smaller than those of S1. Moreover, $\Gamma_{\mathcal{C}_2} \sim O(1)$ almost in the whole range of distances δ_θ for S1, indicating that no trapping is expected in the shallower minima, while it exhibits an almost exponential decrease for S0. An intermediate situation is observed for S2 at large distances. As a result, we see that a true funnel-like structure is markedly present only in the energy landscape of the sequence S1, that was indeed already identified as a good folder by looking at different indicators [9,13]. A more detailed analysis shows that the Arrhenius factor in (7) tends to decrease with δ_θ. It is worth stressing that the entropic factors counterbalance this tendency in S1 for both $\Gamma_{\mathcal{C}}$'s over a wide range of values of δ_θ [see Fig. 8(b)]. A similar effect is observed only for $\Gamma_{\mathcal{C}_1}$ in S2.

V. CONCLUSIONS AND PERSPECTIVES

In this paper, we have analyzed the structure of the energy landscape of a 2D off-lattice model of a polypeptidic chain and found that the relative closeness between neighboring minima can be exploited to implement an effective algorithm to identify directly connected minima. In order to put the analysis on firm quantitative grounds, we have tested several definitions of distance between configurations, finding that the best performances are obtained for the angular distance $\delta_\theta^{(1)}$ [see Eq. (3)]. In fact, the $\delta_\theta^{(1)}$ distances corresponding to DCP are more sharply concentrated at small values than for all other definitions of distances. In particular, we have found that restricting the systematic search to all pairs of minima whose distance is smaller than $\delta_\theta^{(1)\star} = 0.5$, one can recover almost 99% of all DCP [20–22].

The drawback of the approach we are proposing is that a tiny fraction of DCP is unavoidably missed, but the consequences are not *a priori* clear. Since our analysis of the native valley has shown that minima are more rarefied in the
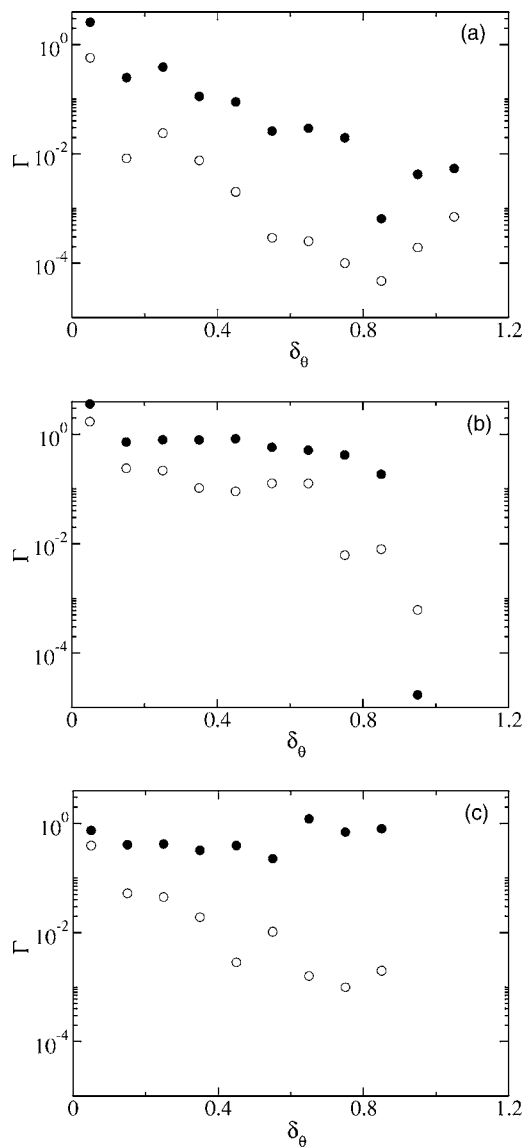
FIG. 8. Escape rates (7) versus the angular distance δ_Θ for each DCP for sequence S0 (a) and S1 (b) and S2 (c) at a temperature $T=T_f$ (namely, $T_f=0.044$ for S0 and S2, $T_f=0.061$ for S1). The filled circles refer to the transition rates $\Gamma_{\mathcal{C}_1}$, while the empty circles to $\Gamma_{\mathcal{C}_2}$.

vicinity of the native configuration (see Fig. 7), we conclude that it is wise to complement the above metric criterion with an extensive search limited to the minima of the native valley. It is now important to verify to what extent such a scenario extends to more realistic 3D systems, where the implementation of effective algorithms to reconstruct the energy landscape is even more crucial than in 2D. Furthermore, important hints about the true relevance of missing links in a connectivity graph will come after imposing a dynamics on the graph itself by adding the activation rates Γ relative to the transitions between directly connected minima. By comparing the resulting evolution with that of the original system, one can in particular determine the minimal fraction of DCP that is necessary to identify for a meaningful reconstruction of the folding process.

Finally, in order to test how the observed rarefied density of minima in the vicinity of the native configuration is connected to the presence of a true funnel-like structure, we have computed the activation rate Γ inside the native valley. It turns out that moving away from the native configuration, while in the homopolymer Γ decreases very rapidly, it stays almost constant in the sequence S1, previously identified as a good folder by other means. This is a clear indication that the homopolymer can be trapped in several minima far from the minimal energy state, while an accessible native valley does exist for S1.

APPENDIX

The database containing the minima of a model sequence can be used to determine the first-order saddles directly connecting neighboring minima. The method we propose to use to accomplish this goal is based on two different algorithms to solve the following two problems: (i) Given two minima $\mathcal{C}_1$ and $\mathcal{C}_2$ and two configurations $\mathcal{P}_1$ and $\mathcal{P}_2$ belonging to their basins of attraction, one wants to determine two configurations $\mathcal{Q}_1$ and $\mathcal{Q}_2$ on the segment $\mathcal{P}_1\mathcal{P}_2$ arbitrarily close to the ridge dividing the two basins and lying on its opposite sides; (ii) given $\mathcal{Q}_1$ and $\mathcal{Q}_2$, as defined above, one wants to apply an iterative procedure, leading the two points to converge on the saddle.

The procedure is here described in more details: (1) Given $\mathcal{Q}_1=\mathcal{P}_1$ and $\mathcal{Q}_2=\mathcal{P}_2$, an intermediate configuration $\mathcal{C}$ is defined by setting its bond angles equal to the average of the corresponding angles of $\mathcal{Q}_1$ and $\mathcal{Q}_2$. (2) A steepest descent procedure is applied to $\mathcal{C}$ until a minimum $\mathcal{C}_m$ is reached. (3a) If $\mathcal{C}_m$ coincides with $\mathcal{C}_1$ ($\mathcal{C}_2$), we replace $\mathcal{Q}_1$ ($\mathcal{Q}_2$) with $\mathcal{C}$ and repeat the step (2) until the Euclidean distance $d_E(\mathcal{Q}_1,\mathcal{Q}_2)$ between $\mathcal{Q}_1$ and $\mathcal{Q}_2$ becomes smaller than a given threshold δ. (3b) If $\mathcal{C}_m$ differs from both $\mathcal{C}_1$ and $\mathcal{C}_2$, we assume that the basins of attraction of the two minima are not directly connected (i.e., $\mathcal{C}_1$ and $\mathcal{C}_2$ are not neighbors) and $\mathcal{C}_m$ is added to the minima database, provided it is not already included. (3c) If $d_E(\mathcal{Q}_1,\mathcal{Q}_2)<\delta$, we conclude they are close enough to the ridge (i.e., the stable manifold separating the basins of attraction of $\mathcal{C}_1$ and $\mathcal{C}_2$) and pass to item (4). (4) We let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ evolve in time according to a steepest descent relaxation until $d_E[\mathcal{Q}_1(t),\mathcal{Q}_2(t)]$ becomes larger than δ and go back to (1), identifying $\mathcal{P}_1$ with $\mathcal{Q}_1(t)$ and $\mathcal{P}_2$ with $\mathcal{Q}_2(t)$. (5) When, during the steepest descent, the gradient of the potential both in $\mathcal{Q}_1$ and $\mathcal{Q}_2$ becomes smaller than another given threshold, we assume $\mathcal{Q}_1$ and $\mathcal{Q}_2$ to be close enough to a first-order

saddle and refine the configuration by means of a Newton's algorithm.

It must be stressed that this algorithm not only allows us to find first-order saddles, but also enriches the number of known minima. The saddle-searching strategy here proposed can then be viewed as an "all purpose" method suitable for a complete exploration of all the features of the energy landscape relevant for protein dynamics.

The data reported in this paper were obtained by applying the algorithm to an initial database of minima determined by the method described in Ref. [10]. It amounts to performing a high-temperature (typically above T_f) Langevin dynamics, when the system is expected to visit a large portion of the accessible phase space. The resulting trajectory is then sampled to pinpoint a series of configurations, which are afterward relaxed according to a steepest descent dynamics and finally refined by means of a standard Newton's method. This procedure was used for identifying a small database of 50–100 collapsed states for each of the three sequences described in Sec. II.

All possible pairs in these initial minima databases were then searched for DCP by means of our saddle-finding algorithm. The new minima found during each run of the algorithm on all possible pairs were stored in the database to be investigated in successive runs. Actually, the number of pairs of minima grows much faster than the number of pairs investigated, thus making impossible a complete analysis. The number of minima and saddles that were identified after three runs is reported in Table I.

In order to perform a complete search of all DCP at least in a restricted set of minima, we have selected all configurations of energy lower than $V_f = V_0 + LT_f$. In this restricted database, all pairs of minima characterized by an angular distance smaller than 0.5 were investigated. The total number of saddles found by this procedure is reported in Table I.

[1] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[2] F. H. Stillinger and T. A. Weber, Science **225**, 983 (1984).
[3] S. Sastry, P. G. Debenedetti, and F. H. Stillinger, Nature (London) **393**, 554 (1998).
[4] L. Angelani, R. Di Leonardo, G. Ruocco, A. Scala, and F. Sciortino, Phys. Rev. Lett. **85**, 5356 (2000).
[5] S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[6] D. A. Evans and D. J. Wales, J. Chem. Phys. **118**, 3891 (2003).
[7] A. Baumketner, J-E. Shea, and Y. Hiwatari, Phys. Rev. E **67**, 011912 (2003).
[8] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[9] A. Torcini, R. Livi, and A. Politi, J. Biol. Phys. **27**, 181 (2001).
[10] L. Bongini, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **68**, 061111 (2003).
[11] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993).
[12] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).

[13] A. Irbäck and F. Potthast, J. Chem. Phys. **103**, 10298 (1995).
[14] A. Irbäck, C. Peterson, and F. Potthast, Phys. Rev. E **55**, 860 (1997).
[15] P. G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, New York, 1979).
[16] T. Veitshans, D. Klimov, and D. Thirumalai, Folding Des. **2**, 1 (1997).
[17] J. S. Langer, Ann. Phys. (N.Y.) **54**, 258 (1969).
[18] P. Hänggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251 (1990).
[19] This is a free parameter, which has to be fixed according to some physical criterion.
[20] L. Holm and C. Sanders, J. Mol. Biol. **233**, 123 (1993).
[21] S. Wallin, J. Farwer, and U. Bastolla, Proteins **50**, 144 (2003).
[22] As reported in Refs. [20,21], not all possible distances are equally effective in catching topological and energetic differences between protein configurations. For instance, measures based on intramolecular distances seem to be inappropriate to discriminate structural modifications [21]. Nonetheless, it will be interesting to test these classes of distances in the framework proposed in this paper.