Thermodynamic features characterizing good and bad folding sequences obtained using a simplified off-lattice protein model

A. Amatori,[1,2] J. Ferkinghoff-Borg,[3] G. Tiana,[1] and R. A. Broglia[1,2]
[1]*Department of Physics, University of Milano and INFN, sezione di Milano, via Celoria 16, 20133 Milano, Italy*
[2]*The Niels Bohr Institute, Blegdamsvej 17, 2100 Copenhagen, Denmark*
[3]*NORDITA, Blegdamsvej 17, 2100 Copenhagen, Denmark*
(Received 12 December 2005; published 8 June 2006)

The thermodynamics of the small SH3 protein domain is studied by means of a simplified model where each beadlike amino acid interacts with the others through a contact potential controlled by a 20×20 random matrix. Good folding sequences, characterized by a low native energy, display three main thermodynamical ensembles, namely, a coil-like ensemble, an unfolded globule, and a folded ensemble (plus two other states, frozen and random coils, populated only at extreme temperatures). Interestingly, the unfolded globule has some regions already structured. Poorly designed sequences, on the other hand, display a wide transition from the random coil to a frozen state. The comparison with the analytic theory of heteropolymers is discussed.

DOI: 10.1103/PhysRevE.73.061905 PACS number(s): 87.15.Aa, 61.25.Hq

I. INTRODUCTION

Since Anfinsen first stated the thermodynamic hypothesis [1] (that is, in a given environment, structural and functional features of proteins are fully encoded in their amino acid sequence), a consistent effort has been made in the study of the relationship between the amino acid sequence and its native structure and function. A significant part of this effort has been dedicated to the so-called inverse folding problem, that is, to the design of sequences which have a desired structure as the unique, stable, kinetically accessible ground state (GS). The simplest approach to this problem is to search for the sequence that minimizes the energy of the system, keeping the native conformation and the ratio between the different kinds of amino acids (composition) fixed [2]. At the basis of this approach lies the assumption that the free energy of most states of the system obeys the principle of self-averaging, so that the total probability of the competing states is unaffected by the design. The property of self-averaging is also an element of the replica method, which complements the description given by the random energy model (REM) for heteropolymers [3].

A more efficient approach, which has given good results on lattice models, is to optimize either the Z score [4] or the approximated free energy of the system [5,6]. However, attempts to apply this idea to continuum hydrophobic-polar models has led to results less satisfactory than expected [7]. Nevertheless, the energy-minimization approach has still the advantage of being simple to implement (especially in continuum space, where the wideness of the conformational space makes the calculation of the free energy nontrivial) and it has proven successful in finding sequences that fold on the crystallographic structure of the SH3 domain within a distance root mean square deviation (dRMSD) of 2.6 Å [8]. There, as in other works [9], a major problem in achieving this goal has been the poor knowledge of the interaction among amino acids. A possible strategy to circumvent this limitation is based on the assumption that the ability of proteins to display a low-entropy equilibrium state at biological temperatures is a consequence of the heterogeneity of the interactions, together with the polymeric geometry of the system [10]. Consequently, the inverse folding approach should work for any quenched random interaction, provided it is sufficiently heterogenous [3] (and thus different from the simple hydrophobic-polar models).

In the present work, we have focused our attention on SH3, a small (60-residue) β-like protein domain (see Fig. 1, left panel) which has been widely investigated both experimentally [11–13] and computationally [14–16]. Our model has proven successful in discriminating between good and bad folding sequences on the basis of only their native state energy [8]. In particular, it has been possible to identify a threshold energy E^c_{targ}, such that sequences with native energy $E_{targ}<E^c_{targ}$ are good folders, while sequences with $E_{targ}>E^c_{targ}$ do not fold to the SH3 native structure and display low-energy conformations very different among themselves.

Good folding sequences display a rather sharp transition between a globular unfolded state and the unique native conformation, as experimentally observed [17]. Microcalorimetric experiments can explore temperatures which typically range from 0 to 90 °C. Experimental measures of the specific heat for four sequences folding to the SH3 domain are shown in the right panel of Fig. 1. All these proteins display a single peak in the specific heat at temperatures ranging from $\simeq$50 to $\simeq$70 °C. The thermodynamics of bad folders is more difficult to study due to their tendency to clump into insoluble aggregates.

In the present work we employ efficient sampling algorithms [18] to study the thermodynamics of good and bad sequences designed on the SH3 fold and compare them to random sequences.

The main issues we want to investigate are the nature of the equilibrium states that the protein populates, as well as the thermodynamics and structural properties of these states. Furthermore, we study the extent to which these properties can be described by the standard theory of heteropolymers [19,20].

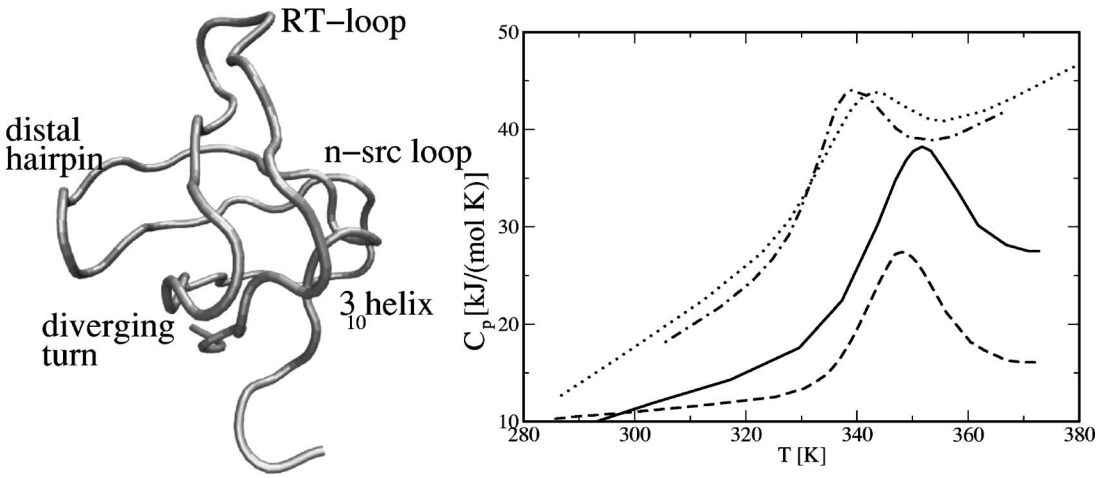

FIG. 1. (Left) The native structure in a C_α representation of SRC SH3 as obtained by crystallographic experiments (Protein Data Bank code 1FMK). On the picture is explicitly indicated the RT loop (residues 8–19), the diverging turn (D_v, residues 20–27), the n-src loop (residues 28–37), the distal hairpin (D_t, residues 38–50), the helix 3_{10} (residues 51–54), and the sheet β_{1-5} (residues 1–7 and 55–57). (Right) the specific heat of four different SH3-domain proteins, obtained from experiments, are shown (solid curve is Btk [39], dashed curve is α-spectrin [11], dotted curve is Abl, and dash-dotted is Fyn [40]).

II. THE MODEL

The model we use has been described in Ref. [8]. It is a reduced off-lattice single-bead model where the amino acids are represented by spherical beads centered around the C^α atom and are connected by an inextensible chain. The energy potential is the sum of pair interactions via a square well function where each σth amino acid type is characterized by a specific value of the hard core radius $R^{HC}(\sigma)$ and of the interaction strenghts $B(\sigma,\pi)$. The matrix $B(\rho,\pi)$ is generated according to a Gaussian distribution with mean $B_0=0.23$ and standard deviation $\sigma_B=0.53$, in arbitrary units. These values of B_0 and σ_B have been chosen to optimize the efficiency of our method. In particular a positive mean induces frustration in the system thus ensuring that the compaction process is selective, i.e., that random compact structures are unlikely [3]. Sequences are designed making use of a Monte Carlo simulation at various temperatures in the space of sequences, i.e., by switching two amino acids at random and then accepting or rejecting the change according to the Metropolis algorithm [21].

Another important ingredient of the model is a constraint on the total number of contacts each residue is allowed to build. This constraint has been introduced because single-bead models oversimplify the geometry of the residues and thus give rise to unphysical conformations where residues build more contacts than their real geometry would allow. Therefore a maximum number of contacts $n_{max}(\sigma)$ has been assigned to the 20 amino acids [8].

The order parameters analyzed are the radius of gyration (R_g), the root mean square deviation (RMSD), and the dRMSD,[1] which performs well in discriminating different states of the system. In all our simulations the RMSD and dRMSD were highly correlated. Consequently we will refer only to the dRMSD. When the dRMSD is used to calculate geometrical differences between a given structure and the native src-SH3 we will use the simbol d_N. When used to calculate differences between any couple of structures, it will be called d_S.

For six particular secondary structures of src-SH3 [that is, the RT loop (residues 8-19; see Fig. 1), the diverging turn (D_v, residues 20-27), the n-src loop (residues 28-37), the distal hairpin (D_t, residues 38-50), the helix 3_{10} (residues 51-54), and the sheet β_{1-5} (residues 1-7 and 55-57)] as well as for their relative conformations, we define a "*structure content*" q as the average fraction of native contacts within each structure.

In what follows, we study the thermodynamic behavior of the nine different sequences summarized in Table I and generated according to the E_{targ} criteria. The canonical averages $\langle A\rangle_\beta$ of the various characteristic quantities A have been calculated for any inverse temperature β according to

$$\langle A\rangle_\beta = \frac{\sum_E g(E)\exp(-\beta E)\langle A\rangle_E}{\sum_E g(E)\exp(-\beta E)}.$$

Here, the density of states $g(E)$ is obtained using the generalized ensemble algorithm described in Ref. [18] coupled with a multicanonical weight scheme [22]. The microcanonical averages $\langle A\rangle_E$ have been estimated from the arithmetic average obtained from the sampling at each energy. Specific heat profiles are found from the energy variance, i.e., $C(\beta)=\beta^2(\langle E^2\rangle_\beta-\langle E\rangle_\beta^2)$. Sampling convergence has been ensured by inspecting the overall flatness of the histograms produced by the algorithm. In all cases, the relative error (estimated from independent runs) is of the order of a few percent for all temperatures of interest.

[1] dRMSD is defined as the root of the mean square difference between the interresidue distance in the given conformation and in the native state, calculated over all pairs of residues.

TABLE I. Sequences with selected energies E_{targ} on the SH3 target conformation displayed in Fig. 1. E_{gs} is the energy of the ground state of the sequence. Sequences (s_1,s_2,s_3), (s_4,s_5,s_6), and (s_7,s_8,s_9) are good, bad, and randomly generated folders, respectively.

Label	E_{targ}	E_{gs}	Sequence
s_1	−37.80	−46.96	GLLLLAANNWWVTRTDEEKKDYVSSSSDDTQTGGYNIEGLIFFRQVVPPEAHTYYSSSTT
s_2	−35.53	−45.03	QQHAASSSDDSDVFTVPPLGNLTNYYGIITKTTWLLFEGGAYTRNVDEEESSTLSVKYRW
s_3	−34.85	−44.92	GDSAAAHQPERWWTTSSSEEPIYEVLLNVTTTFTRDVDSSDKVGFNGLLLQGTIYYNSKY
s_4	−34.30	−44.52	QWAAHEEEDYRNFGTSSSYQGPGINSSFKTGYTTVDSDSLATRVVVDLLLILWEPKNYTT
s_5	−33.65	−45.02	SGLNLEEPGKKYFRRTAAWFVEGSDSSVGTTTTNQHQTALLLWVSDDYYYIIVEPDSSTN
s_6	−23.67	−42.28	DSSSSEERDIFYTTTWYYQQGPLNSLLLGTVKTVDDIYSSAKTRWVGAAHGPTEEFNLVN
s_7	−4.52	−42.36	NLILYEKLDNRFNKWWFLADSSPASGQVDRTTSTVSSTQEHTTYEEYVSGLGTIPDAVGY
s_8	+5.36	−38.72	LWYSSLEGGRVNLDTSSKVTPILSFAQGTDRVDDQEYTGIYWTVTEHTAEKYFNPNSALS
s_9	+8.26	−40.92	EYLSVIKTEDPKQSEYPSWLSEFFLLTIATGNTLYYDGVHAVTSSRNSGGDAVRNDTTWQ

III. THERMODYNAMICS OF GOOD FOLDERS

Sequences s_1, s_2, and s_3 (see Table I) display proteinlike properties [8], having a unique and stable native state corresponding to the SH3 conformation shown in Fig. 1 and being able to reach it in a short time. The conformational specific heat $C_p(T)$ of these sequences, calculated with the present model, is displayed in Fig. 2. An interesting feature is that the three sequences, although having less than 10% identity, display very similar specific heats. C_p is in all cases characterized by four peaks, which mark the transition between different states. Because of these similarities, in the following we refer to the behavior of sequence s_1 as a template for all good folders, unless otherwise mentioned. To identify the features of the thermodynamical states of this sequence, we have plotted in Fig. 3 the averages of d_N, R_g, and of the energy E as functions of temperature. At high temperatures ($T>0.6$, the corresponding thermodynamical state being marked as V) the chain behaves as a random coil with an average energy only slightly below zero. In this state the protein has very few (<5) contacts, no detectable secondary structures, a mean gyration radius $\bar{R}_g \simeq 22$ Å, and a d_N between 18 and 25 Å. That is, it does not have anything in common with the native conformation.

Decreasing the temperature, the system shows a low, wide, peak in $C_p(T)$. The state lying beyond the peak (state IV in Fig. 3) is still extended, having a mean radius of gyration of 18 Å. The associated conformations are overall dissimilar from the native one, with $\bar{d}_N=15\pm4$ Å and $\bar{d}_S=9\pm2$ Å. The equilibrium distribution of d_S at $T=0.5$ is shown with a dashed curve in the upper panel of Fig. 5 below and indicates a wide structural heterogeneity. Nonetheless, state IV has a sizable content of structured distal hairpin and RT loop (q between 0.25 and 0.56), while the n-src loop, the diverging turn, the helix 3_{10}, and the sheet β_{1-5} are essentially absent (see Table II). These conformations are characterized mainly by local bonds, although there are few nonlocal native contacts of residues 2, 3, and 4 with residues 24, 25, and 26, which give rise to the RT loop.

At a temperature $T\simeq0.34$, the system undergoes a marked decrease in the energy and a sharp compaction ($\bar{R}_g$ decreases

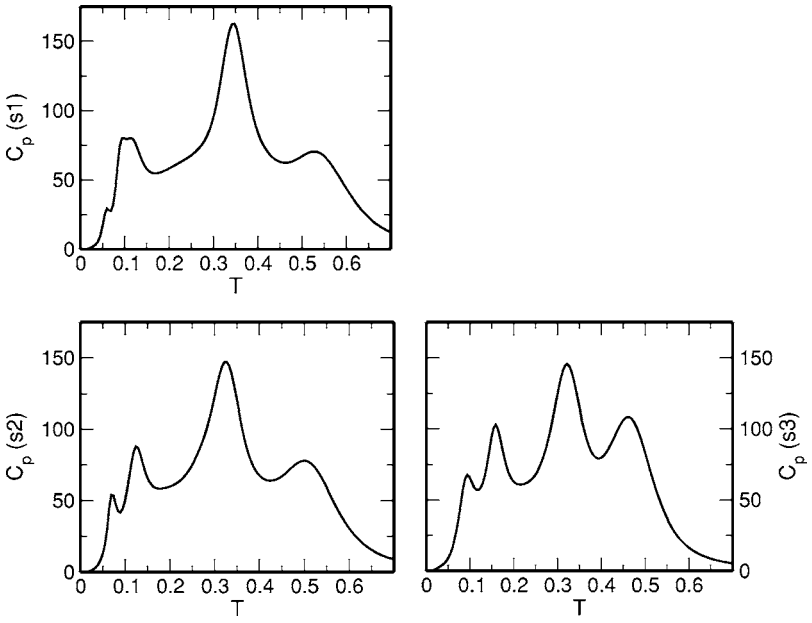

FIG. 2. The specific heat $C_p(T)$ of the three good folders for sequence s_1 (top), sequence s_2 (bottom left), and sequence s_3 (bottom right).

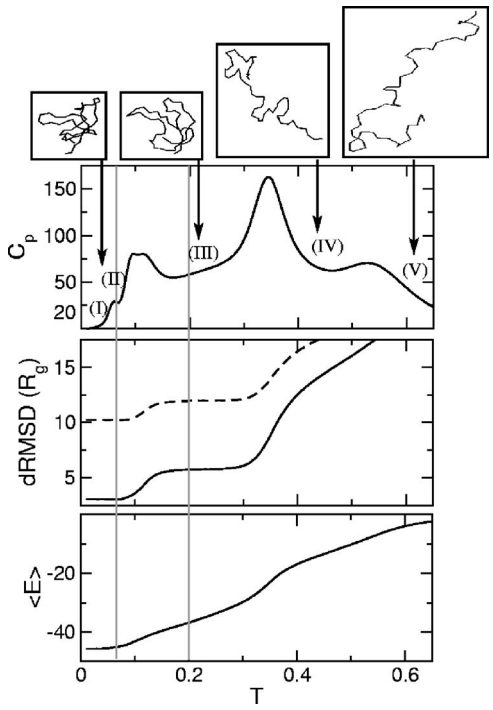

FIG. 3. The specific heat $C_p(T)$ (top panel), average d_N expressed in Å (middle panel, solid curve), average radius of gyration $\langle R_g \rangle$ (Å, middle panel, dashed curve) and average energy $\langle E \rangle$ (bottom panel) for sequence s_1. The interval of temperatures defined by gray perpendicular lines marks the region identified as biologically relevant. The four pictures on the top show typical conformation of the system in each state (the first picture represents both states I and II, as conformational differences are negligible on this scale).

from 18 to 12 Å). The state beyond this coil-globule transition (state III of Fig. 3) is associated with conformations displaying an average d_N=6 Å, still dissimilar from the native conformation, thus qualifying as an unfolded state. This transition is underlined by a rather sharp peak in C_p, consistently with a first-order coil- (ordered-)globule transition, as predicted by the theory of nonrandom heteropolymers [19]. The distribution of d_S for state III ($T \simeq 0.2$) is shown in Fig. 5 (dotted curve). Although the major peak is still centered at $d_S \simeq 8$ Å, indicating the structural heterogeneity typical of the unfolded state, a small peak emerges at d_S=4 Å, which indicates a small presence of specific conformations. On the other hand in this region the specific heat is well above zero, and the energy displays values in the range $E \simeq -29$ to $E \simeq -39$. Interestingly, this energy interval is not accompanied by any major structural changes, as indicated by the approximately constant values of R_G and d_N.

Analyzing the conformations and the map of contacts of state III, we observe that it is characterized by the RT loop and the D_t essentially fully formed (with probability 0.83 and 1.0, respectively), and, consequently, the sheet between strands β_3 (36-41) and β_4 (47-51). Also the diverging turn D_v is well structured at this stage (q=0.75), while the two terminals get together and give rise to a shorter sheet (which is generally not fully formed at this point; q=0.55) between strands β_1 and β_5. The poor presence of the n-src loop (q =0.40) causes the sheet β_2 (24-28)–β_3 not to be formed. Within this context one can see from Table III as well as Fig. 4 that the contacts between D_t and D_v are not formed. On the other hand, one finds that the contacts between the RT loop and the distal hairpin are formed with high probability (0.86), contacts which were not formed in state IV. The presence of the ensemble described above, dominated by unfolded globular conformations displaying structured fragments, results to be a thermodynamic hallmark of good folders.

Going back to Fig. 3, a third peak is found at temperature T=0.10 and marks a transition to a state with average d_N =3.0 Å and RMSD=4.6±0.8 Å, which can be regarded as the native state (cf. Ref. [8]). This transition gives rise to the formation of the n-src loop, which is associated with a large entropy loss. Moreover, at this temperature the system undergoes the formation of the helix (q_{II}=0.75) and an increase in the content of sheet β_{1-5} (q=0.85; cf. Table II) which, nontheless, is still able to fluctuate. The average radius of gyration of state II is $\bar{R}_g$=10 Å, which is the same as that of the crystallographic native conformation and is only 20% more compact than state III. The distribution of d_S is now peaked around 2.5 Å (see Fig. 5), in accordance with the features of uniqueness of the native state. These data suggest that the peak at T=0.10 is the one observed in calorimetry experiments (see Fig. 1), associated with the folding transition. The temperature region corresponding to the experimental window of Fig. 3 is indicated with a gray frame. This correspondence is guided from the shape of the peak of the specific heat only, rather than from an implicit assumption of a specific functional relation between the temperature of the simulation and the experiment. We deliberately exclude the hypothetical temperature-induced coil-globule transition peak in the comparison, since for most domains the experimentally thermal denatured state has significantly more residual structure than what would be expected from a random coil state [23]. We will return to this point in the Sec. V.

Finally, at temperature T=0.06 the specific heat of sequence s_1 has a last, small peak and then sharply drops to zero, in correspondence with the freezing of the system into its ground state. The contact map of state I indicates that the only essential structural difference from the previous state is a tightened sheet β_{1-5} (q=0.90), that freezes the last degrees of freedom of the system (cf. Table II).

IV. THERMODYNAMICS OF BAD AND RANDOM SEQUENCES

For comparison with the good folding sequences, we next analyze the specific heat of three randomly generated sequences (sequences s_7, s_8, and s_9) and of three bad folders (s_4, s_5, and s_6), that is, sequences designed to have, on the SH3 native conformation, an energy too high to fold (i.e., larger than E_c [8]). The plots of the specific heat are displayed in Figs. 6 and 7, respectively.

In all these cases the pattern common to good folders is lost and the specific heats have a more sequence-dependent shape. The shape of C_p for bad folders is characterized by a large shoulder which involves all temperatures up to T=0.6,

TABLE II. The average energies and structural features are here summarized for the three good folders at the five thermodynamically relevant states. From left to right, columns show the average value of the energy, of the dRMSD from the native state (d_N), of the radius of gyration (R_g) and the structure content q of six secondary structures of SH3 [that is, the RT loop (RT), the distal *hairpin* (D_t), the *diverging turn* (D_v), the n-src loop (n-src), the helix 3_{10}, and the sheet β_{1-5}].

Label	$\bar{E}$	$\bar{d}_N$	$\bar{R}_g$	q(RT)	$q(D_v)$	q(n-src)	$q(D_t)$	$q(3_{10})$	$q(\beta_{1-5})$
				State V: coil					
s_1	−2	21.0	22.5	0.08	0.00	0.05	0.04	0.04	0.00
s_2	−2	20.9	22.4	0.05	0.03	0.05	0.03	0.02	0.00
s_3	−2	21.3	22.7	0.05	0.03	0.03	0.05	0.00	0.00
				State IV: embryo					
s_1	−10	14.8	18.2	0.25	0.00	0.20	0.25	0.06	0.00
s_2	−13	14.4	17.6	0.45	0.00	0.00	0.50	0.15	0.00
s_3	−14	13.0	16.9	0.33	0.17	0.22	0.56	0.00	0.00
				State III: globule					
s_1	−37	5.7	12.0	0.83	0.75	0.40	1.00	0.12	0.55
s_2	−35	5.9	12.0	0.85	0.66	0.20	1.00	0.15	0.50
s_3	−32	6.2	12.2	0.80	0.70	0.13	0.93	0.42	0.40
				State II: folded					
s_1	−45	3.0	10.2	0.92	1.00	1.00	1.00	0.75	0.85
s_2	−42	3.6	10.2	0.85	0.90	0.70	1.00	0.45	0.78
s_3	−41	3.9	10.3	0.95	1.00	0.85	1.00	0.50	0.70
				State I: frozen					
s_1	−47	3.0	10.2	0.95	1.00	1.00	1.00	0.75	0.90
s_2	−45	3.6	10.0	0.95	0.92	0.88	1.00	0.66	0.88
s_3	−45	4.0	10.1	1.00	1.00	0.85	1.00	0.45	0.70

with peaks superimposed in a disordered fashion. On the other hand, random sequences display a more compact C_p, being significantly different from zero only in the region between $T=0.05$ and 0.40. The peaks in the specific heat of bad and random sequences are tightly connected with the variation of the radius of gyration of the protein. In Fig. 8 the average radii of gyration $\bar{R}_g$ for a folding sequence (s_1, solid curve), for a bad sequence (s_4, dashed curve), and for a random sequence (s_7, open circles) are shown.

A. Random sequences

The radius of gyration of the random sequence displays a wide sigmoidal shape which spans the region of the major peak in the specific heat. Consequently, this wide peak is associated with a broad transition from a coil ($\bar{R}_g \simeq 20$ Å) to a compact globular macro-state ($\bar{R}_g \simeq 10$ Å). The range of temperatures that can be interpreted as biologically relevant (cf. Fig. 3) partially overlaps the peak in C_p. According to Flory's model of homopolymer collapse [24,25], the volumetric interaction free energy takes the form

$$F_{vol}(T,\phi) \simeq \mathcal{N}T\left(1-\chi\phi+\frac{1-\phi}{\phi}\ln(1-\phi)\right),$$

where $\phi=\nu\mathcal{N}/V$ is the average polymer volume fraction, $\mathcal{N}$ the total number of monomers in the chain, ν the excluded volume of each monomer, and V the average volume occupied by the chain, $V \simeq (\frac{5}{3})^{3/2}\frac{4\pi}{3}R_g^3$.[2] From the maximally dense globule obtained from simulations, $R_g \simeq 9.8$ Å, we obtain $\nu \simeq 141$ Å^3 as the average excluded volume. The Flory-Huggins constant χ can be estimated from the number of contacts in the dense globule, N_C, by noting that $\mathcal{N}\phi$ is the number of binary collisions in the mean-field picture; hence $E=-\mathcal{N}T\chi\phi$ is the total interaction energy. In the high-density limit ($\phi\to 1$), one has $E=-N_C B'$, where B' is the effective interaction energy between monomer pairs. Thus, $\chi \simeq -\frac{N_C}{\mathcal{N}}\frac{B'}{T}=-\frac{z}{2}\frac{B'}{T}$, and z is the coordination number for the (dense) globule. In the case of a homopolymer, $B'=B_0$, where B_0 is simply the monomor-monomer interaction strength. In the heteropolymeric case, the effective interaction is modified according to $B'=B_0-\sigma_B^2/2T$ [19], where B_0 now denotes the average of the interaction matrix, $B_0=\Sigma_{\alpha\beta}p_\alpha p_\beta B_{\alpha\beta}$, and $\sigma_B^2=\Sigma_{\alpha\beta}p_\alpha p_\beta(B_{\alpha\beta}-B_0)^2$ is the variance. Here, the summation $\Sigma_{\alpha\beta}$ runs over the different monomer types α and β, and p_α is the frequency of occurrence of monomer type α. In our case, $B_0=0.23$, $\sigma_B=0.53$, and $z \simeq 2.2$. Combining the various expressions above and expand-

[2]The radius of a maximally compact spheric protein ($\phi=1$) is $R=[(3/4\pi)\mathcal{N}\nu]^{1/3}$, whereas the volume V_g obtained from the radius of gyration is $V_g=(4\pi/3)R_g^3=(5/3)^{-3/2}(4\pi/3)R^3$. In order to find the correct result in the limit $\phi\to 1$, we therefore calculate the average volume as $V=(5/3)^{3/2}V_g$.

TABLE III. The structure content q associated with contacts between SH3 structures of sequence s_1 are reported for the different thermodynamical states.

	β_{1-5}	RT	D_v	n-src	D_t	3_{10}
			Frozen state (I)			
β_{1-5}			0.86	1		
RT			1		1	
D_v				1	1	
n-src					1	1
D_t						
			Native state (II)			
β_{1-5}			0.79	0.5		
RT			1		1	
D_v				0	1	
n-src					1	1
D_t						
			Unfolded globule state (III)			
β_{1-5}			0.71	0.5		
RT			0		0.86	
D_v				0	0	
n-src					1	1
D_t						
			Embryo globule state (IV)			
β_{1-5}			0.56	0		
RT			0		0	
D_v				0	0	
n-src					0	0
D_t						

ing the total interaction free energy for a random heteropolymer to third order in the density we obtain

$$F_{vol}(T,\phi) = \mathcal{N}T\left[-\frac{z}{2}\left(\frac{\sigma_B^2}{2T^2} - \frac{B_0}{T}\right)\phi + \frac{1}{2}\phi + 1/6\phi^2 + O(\phi^3)\right],$$

with the corresponding second and third virial coefficients $b(T)=\frac{\nu}{2T}\left(T - z\frac{\sigma_B^2}{2T} + zB_0\right)$ and $c=\nu^2/6$, respectively. The theory predicts a second-order coil-globule transition at temperature θ where $b(\theta)=0$. In the present case, $\theta \simeq 0.36$, which is in the high end of the transition region. However, according to the standard Lifshitz theory of the coil-globule transition [20] there is an entropy cost associated with the surface formation of the globule, because the chain sections on the surface layer necessarily have the form of loops. Since this entropy cost scales with the system size as $\mathcal{N}^{2/3}$, the transition temperature T_{cg} may be shifted to a value somewhat below the thermodynamic θ point for small systems. By balancing the energy gain from the coil collapse with the entropy loss of the surface formation the theory predicts the relative shift of the transition temperature $\tau_{cg} \equiv \frac{T_{cg}-\theta}{\theta}$ to be [20]

$$\tau_{cg} \simeq -2.7a^{3/2}c^{1/4}b_\theta^{-1}\mathcal{N}^{-1/2}, \quad (1)$$

where a is the Kuhn length and b_θ is defined from the Taylor expansion of $b(\tau)$ around the θ point, $b(\tau)=b_\theta\tau + O(\tau^2)$ and

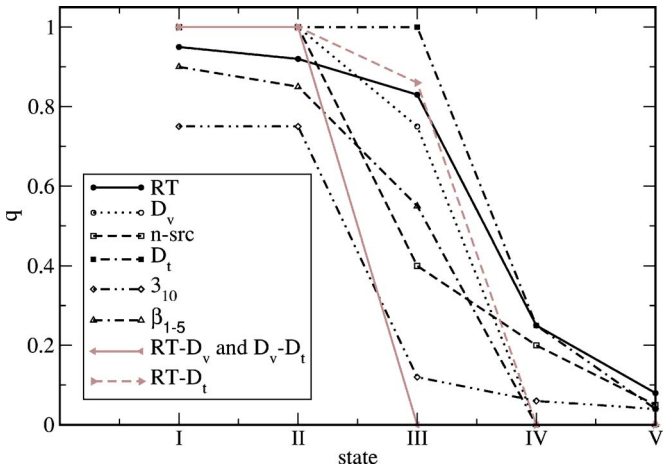

FIG. 4. (Color online) The structure content q of the motifs of SH3 and between some of them (see Tables II and III) of sequence s_1 in the different thermodynamical states.

$\tau=\frac{T-\theta}{\theta}$. For the present model, we obtain $b_\theta=\nu\left(\frac{1}{2}+\frac{z\sigma_B^2}{4\theta^2}\right)$. To determine a, we assume for simplicity that the chain behaves as a Gaussian coil at the point where $\frac{dR_g}{dT}$ has its maximum. From the relation $R_g^2=\frac{1}{6}al\mathcal{N}$, where $l\simeq 3.8$ Å is the distance between consecutive monomers, one obtains $a\simeq 7.6$ Å, corresponding to a moderately flexible chain, $v/a^3\simeq 0.32$. Inserting the expression for b_θ and c in Eq. (1) gives

$$\tau_{cg} \simeq -2.7\times 6^{-1/4}\left[\frac{1}{2}+\frac{z}{4}\left(\frac{\sigma_B}{\theta}\right)^2\right]^{-1}\left(\frac{a^3}{\mathcal{N}\nu}\right)^{1/2} \simeq -0.23,$$

corresponding to $T_{cg}\simeq 0.27$, which is in excellent agreement with the observed midpoint of the transition (cf. Fig. 6). The coil-globule transition happens, coherently with the theory of random heteropolymers [19], at approximately the same temperature ($T\simeq 0.25$) for all the random folders.

From the lack of any plateau in $\bar{R}_g$ (see Fig. 8), we see that the system is not populating any well-defined globular state in the transition from the coil to the frozen ground state. In other words, differently from good folders, the freezing transition is not immediately distinguishable from the coil-globule transition region.

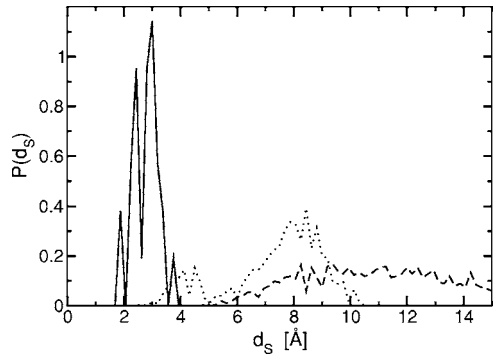

FIG. 5. The distribution of dRMSD for pairs of conformations (d_S) associated with sequence s_1 at $T=0.10$ (solid curve), 0.20 (dotted curve), and 0.50 (dashed curve).

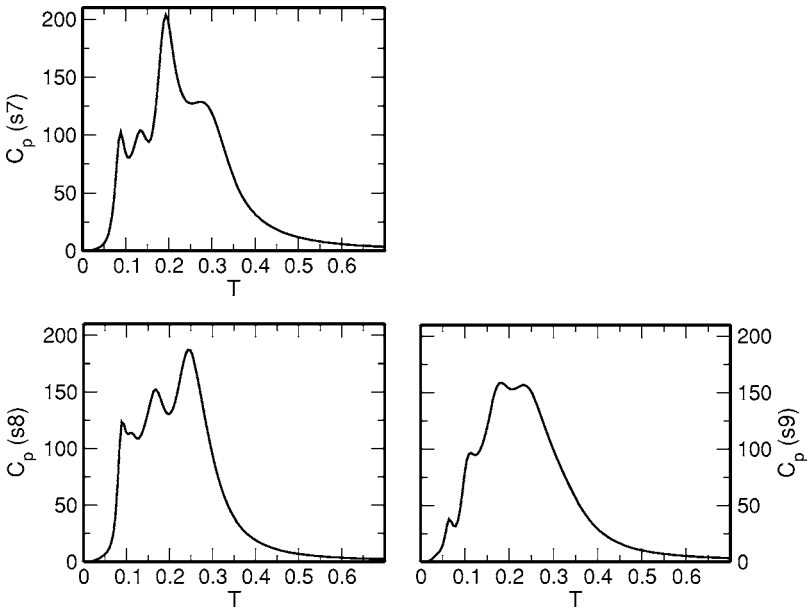

FIG. 6. The specific heat $C_p(T)$ of the three random folders; sequence s_7 (top), sequence s_8 (bottom right), and sequence s_9 (bottom left).

Making use of the common picture of freezing [19] as the process which brings random heteropolymers from the sea of random globular configurations into their ground-state basin, we calculate the thermodynamical average of d_S as a function of temperarature. The plot of $\bar{d}_S(T)$ [see Fig. 9(a)] shows indeed a marked transition for all the sequences from values $\bar{d}_S \approx 1.5$ Å (corresponding to mutually similar structures) to $\bar{d}_S \approx 4.2$ Å (corresponding to compact structures with no similarity). The average temperature of this freezing transition is $T_{freeze}=0.1$ [see Fig. 9(a)] and coincides with a sharp decrease of $C_p(T)$. In accordance with REM [19], which predicts the freezing transition to be of second order, we may regard this decrease of C_p as a signature of a second-order phase transition in our finite system.

To investigate the roughness of the energy landscape, we calculate the value of d_S between low-energy states, chosen within the 0.15 fractile of the energy distribution at $T=T_{freeze}$ [i.e., $P_T(E<E_{15\%})=0.15$, at $T=T_{freeze}$]. As Fig. 9(b) shows, the set of these low-energy ($E_{15\%}=-39$) states for random sequences is very heterogeneous ($\bar{d}_S=4.4$ Å, and null probability for $d_S<2.7$ Å). Conversely, kinetic simulations below the calculated freezing temperature ($T<0.1$), initialized in any one of these low-energy conformations, visit states with a pairwise distance $d_S<2.5$ Å [see Fig. 9(c)]. This shows that conformations with $d_S>2.7$ Å are typically separated by consistent energy barriers which make them kinetically inaccessible to each other even at temperatures where the specific heat is well above zero. This is fully consistent with the thermodynamics of random heteropolymers [3], which predicts a free-energy landscape at low temperature with several wells, each well containing conformations mutually similar, and different wells containing conformations with little similarity. The low-temperature state of the random sequence is then, according to the language of Ref. [3], a frozen state.

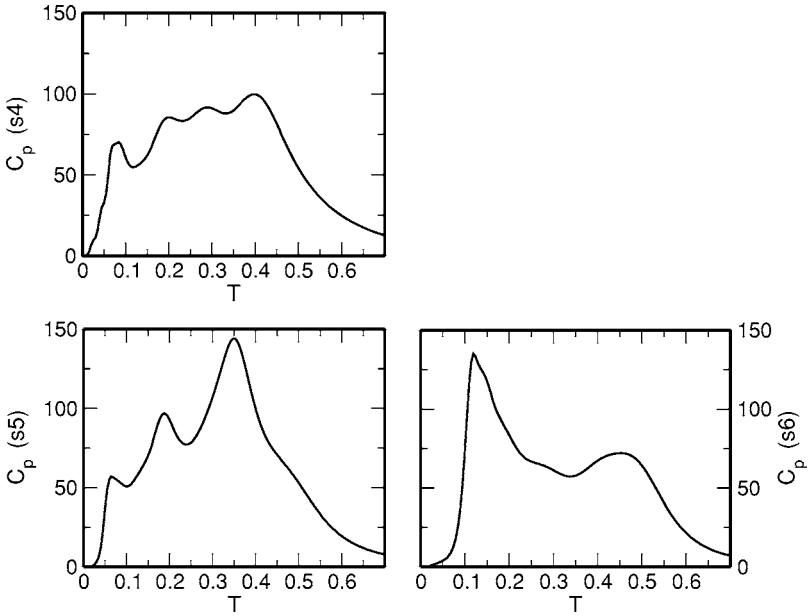

FIG. 7. The specific heat $C_p(T)$ of the three bad folders; sequence s_4 (top), sequence s_5 (bottom right), and sequence s_6 (bottom left).

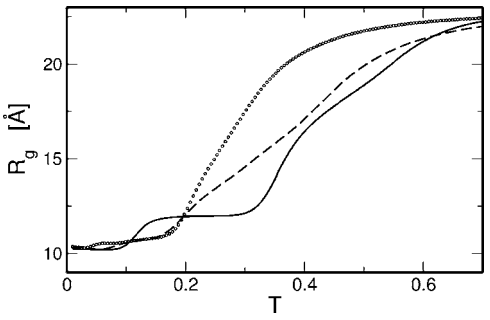

FIG. 8. Radius of gyration as function of temperature for three sequences representative of the three groups; a good folder (sequence s_1, continous solid curve), a bad folder (sequence s_4, dashed curve), and a random sequence (sequence s_7, open dots).

The theory of heteropolymers predicts the freezing temperature of large globules to be $T_{freeze}=\frac{\sigma_B}{2\sqrt{\Delta s}}$ [3], where Δs is the entropy per contact lost in the freezing. Using as an estimate $\Delta s \simeq \ln(a^3/\nu) \simeq 1.14$ [19], one obtains $T_{freeze}^{theor} \simeq 0.25$, which is significantly higher than the value estimated from the simulations. Interestingly, the crude approximation $\Delta s \simeq \ln(a^3/\nu)$ is surprisingly close to the observed value $\Delta s = 1.01$.[3]

Importantly, the observed T_{freeze} is also low compared with the critical design temperature of our model ($T_{design}^{cr} \simeq 0.15$), in marked contradiction with the prediction of REM, where $T_{design}^{cr}=T_{freeze}$. The failure of the equations $T_{freeze}=\frac{\sigma_B}{2\sqrt{\Delta s}}$ and $T_{design}^{cr}=T_{freeze}$ shows that, in our system, low-energy states are not uncorrelated, implying that the principle of self-averaging does not in general apply.

From the structural point of view, random sequence display surprising features. In Table IV the features of the random sequences are reported for their three thermodynamically relevant states. As far as the formation of secondary structures is concerned, random sequences are obviously less effective than good folders. In particular, the helix and the sheet β_{1-5} are practically absent at all temperatures in random folders. Nonetheless, in most of them, the presence of the other peculiar structures (i.e., the RT loop, the Diverging turn, the n-src loop, and the distal hairpin) is non-negligible in low-energy states. This highlights the fact that these structures, primarily based on local and mid-range bonds, pay a lower entropy and hence require less design accuracy to be formed. In other words, there is a structural imprint based only on the length of the chain which favors turns in specific places, and the evolutionary optimization of good sequences can take advantage of such imprint to increase its stability.

[3]The use of σ_B as the standard deviation per contact of the energy probability distribution assumes each contact to be an independent random variable, which can only be expected to hold in the flexible chain limit $\nu/a^3 \to 1$. Assuming the actual number of uncorrelated contacts ($\tilde{N}_c$) to depend on the Kuhn length as $\tilde{N}_c=N_c l/a$, implies that the standard deviation of the energy probability distribution is reduced according to $\sigma_B \to \sigma_B\sqrt{l/a}$. In fact, this simple rescaling is in good agreement with the observed standard deviation of the energy probability distribution around $T=T_{freeze}$ (data not shown) and reduces consistently the predicted freezing temperature ($T_{freeze}^{rescaled} \simeq 0.18$). Nevertheless the freezing temperature obtained from this rescaling is still significantly higher than the observed T_{freeze}, indicating residual correlations between different states.

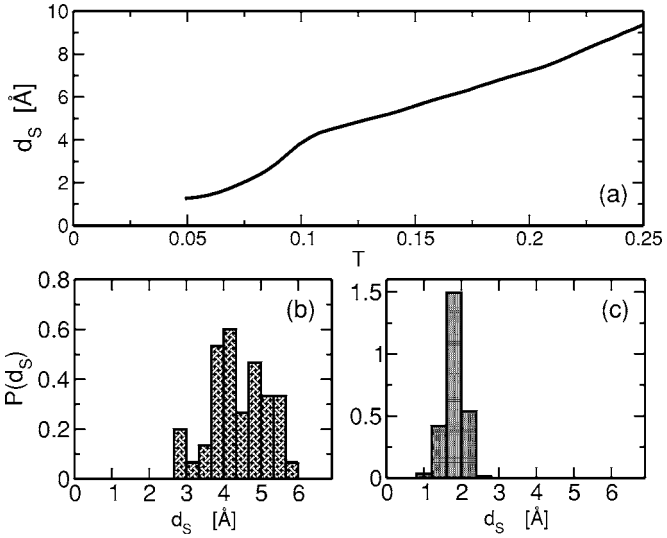

FIG. 9. (a) $d_S(T)$ plot for s_7 (representative of random sequences) highlights the freezing transition at temperature ≈ 0.1. (b) Distribution of d_S for a set of low-energy ($E<-39$) conformations for s_7. (c) Distribution of d_S for states sampled in a low-temperature ($T=0.07$) kinetic simulation starting from one of these low-energy states of sequence s_7.

B. Bad sequences

While the rationale for the specific heat of good folders resides in their common folding properties and that for the random sequences in their average properties, bad folders seem to have lost both of them. Their specific heat features neither the specificity of their sequence nor the averaging of the contribution of uncorrelated residues. Even the coil-globule transition is overwhelmed by other effects, and is not easily detectable in the plot of C_p (see Fig. 7).

The $\bar{R}_g(T)$ plots of these sequences show a rather smooth decrease (as temperature decreases) from random coil values above 20 Å to a ground-state value $\simeq 10.5$ Å [see Fig. 8 where the $\bar{R}_g(T)$ plot of sequence s_5 is shown]. This implies that, similar to random sequences, bad folders do not have the globular unfolded state III typical of good sequences. A qualitatively similar behavior is also found for d_N plots (data not shown).

Unlike random sequences, their low-energy states do not populate a multitude of minima. In fact, thermodynamical samplings of the kind used for good and random sequences, starting from a random conformation and lasting for 50×10^9 steps, find a non-native ($\bar{d}_N=5.0$ Å) but still structurally homogeneous ($\bar{d}_S=3.1$ Å) basin, the bottom of which displays an energy $E \simeq -44$. Conversely, fixed temperature Monte Carlo (MC) simulations at $T=0.1$ starting from the crystallographic conformation of SH3 display another homogeneous ($\bar{d}_S=2.4$ Å) and maximum of $d_S=4.1$ Å) basin with practically the same energy ($E_{min}=-44.5$) but strongly dis-

TABLE IV. The average energies and structural features are displayed for the three random sequences at the three thermodynamically relevant states. From left to right, the different columns show the average value of the energy, of the dRMSD from the native state (d_N), of the radius of gyration (R_g), and the structural content q of six secondary structures of SH3 [namely, the RT loop (RT), the distal hairpin (D_t), the diverging turn (D_v), the n-src loop (n-src), the helix 3_{10}, and the sheet β_{1-5}].

Label	$\bar{E}$	$\bar{d}_N$	$\bar{R}_g$	q(RT)	$q(D_v)$	q(n-src)	$q(D_t)$	$q(3_{10})$	$q(\beta_{1-5})$
					Coil state				
s_7	−2	21.2	22.3	0.05	0.02	0.00	0.00	0.06	0.00
s_8	−1	21.9	22.6	0.02	0.00	0.07	0.00	0.00	0.00
s_9	−2	21.5	22.6	0.03	0.08	0.02	0.03	0.02	0.00
					Globule state				
s_7	−37	5.6	10.5	0.32	0.15	0.45	0.38	0.15	0.11
s_8	−35	6.3	10.0	0.15	0.22	0.12	0.20	0.10	0.06
s_9	−34	6.0	9.6	0.40	0.27	0.53	0.07	0.08	0.05
					Frozen state				
s_7	−42	5.4	10.3	0.14	0.22	0.20	0.45	0.08	0.14
s_8	−39	5.9	9.8	0.22	0.25	0.05	0.30	0.03	0.05
s_9	−41	5.9	9.3	0.44	0.38	0.72	0.03	0.12	0.08

similar from the first one ($\bar{d}_S$=4.7 Å with minimum value of d_S=3.6 Å). The native conformation belongs to this second basin (whose $\bar{d}_N$=3.2 Å), i.e., the native basin is a local minimum of the free energy, at least for temperatures up to 0.1 (where the C_p is still consistently above zero). But while performing the simulation at T=0.3 the system leaves the native state almost immediately, reaches states belonging to the non-native low-energy state, and is never able to return to the native basin within the 5×10^9 steps of the MC sampling.

Our generalized-weight sampling algorithm does not implement an order parameter capable of differentiating between the two basins (which span the same range of energies) and thus is not able to provide their respective free energies. However, the above results show that, at not-too-low temperatures, there is not a large barrier separating the native from the non-native basin; hence the entropy of the non-native state must be markedly larger than that of the native, while their energies are similar.

Unlike good folders [which compensate the low entropy of the native state by optimizing the sequence in such a way that E(native basin)$<E$(random globule)], the picture that emerges here is that the native state of bad folders is still entropically disfavored, but does not have an energetic advantage [E(native basin)$\simeq E$(random globule)] to counterbalance the entropy of competing states.

The alternative ground state of bad folders does not satisfy the conditions of kinetical accessibility and thermodynamical stability required in a protein's native state. In fact, kinetic simulations at fixed temperatures ranging between T=0.1 and 0.2 reach this state only about four out of ten times, while the other six times get trapped into local energy minima (rough energy landscape). Furthermore, the plot of dRMSD$_{GS}$ vs T (Fig. 10) shows that this ground state is not even thermodynamically stable. Indeed dRMSD$_{GS}$ has a smooth increase with the temperature, i.e., there is no energetic barrier segregating the ground state from other higher-energy states.

Of course, these results are not able to exclude the existence of other basins, although they have not been observed in very long simulations for each of the three bad sequences.

Table V summarizes the features of the bad folders in the three themodynamically relevant states. The analysis of secondary structure formation confirms what was stated when random and good sequences were compared. Indeed the RT loop, the distal loop, and, to a lesser extent, the n-src loop and the diverging turn have a non-negligible presence in both the frozen and the globular state of all bad sequences. Conversely, the helix and the sheet β_{1-5} have a comparatively low presence in all sequences at any temperature. Remarkably, the values of q in low-energy states of bad folders reflect the hierarchy of the same structures in good folders (see Table II). That is, the RT loop and the distal loop, with an average value of q at the frozen state of bad folders of, respectively, 0.59 and 0.70, are the first structures to be formed in good folders (e.g., q_{good} is 0.34 and 0.44 at state IV for them), while the n-src loop, the diverging turn (with

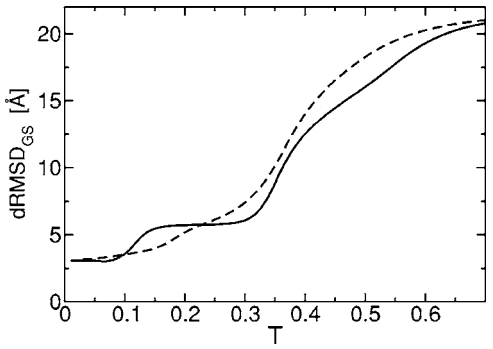

FIG. 10. Variation with temperature of the dRMSD calculated with respect to the ground-state structure (dRMSD$_{GS}$) for a good folder (sequence s_1, continuous curve) and a bad folder (sequence s_5, dashed curve).

TABLE V. The average energies and structural features are here summarized for the three bad folders at the three thermodynamically relevant states. From left to right, columns show the average value of the energy, of the dRMSD from the native state (d_N), of the radius of gyration (R_g), and the structural content q of six secondary structures of SH3 [namely, the RT loop (RT), the distal loop (D_t), the diverging turn (D_v), the n-src loop (n-src), the helix 3_{10}, and the sheet β_{1-5}].

Label	$\bar{E}$	$\bar{d}_N$	$\bar{R}_g$	q(RT)	$q(D_v)$	q(n-src)	$q(D_t)$	$q(3_{10})$	$q(\beta_{1-5})$
					Coil state				
s_4	−2	20.9	22.3	0.10	0.05	0.00	0.12	0.00	0.00
s_5	−1	21.3	22.6	0.07	0.05	0.05	0.15	0.00	0.00
s_6	−1	21.3	22.6	0.08	0.03	0.03	0.10	0.00	0.00
					Globule state				
s_4	−36	4.2	10.9	0.70	0.55	0.45	0.77	0.00	0.04
s_5	−37	5.4	11.2	0.75	0.35	0.10	0.82	0.08	0.07
s_6	−33	6.2	10.7	0.45	0.14	0.33	0.56	0.12	0.10
					Frozen state				
s_4	−45	4.5	10.3	0.67	0.25	0.20	0.80	0.07	0.04
s_5	−45	5.2	10.6	0.70	0.30	0.10	0.80	0.12	0.15
s_6	−42	6.0	10.6	0.40	0.20	0.30	0.50	0.15	0.18

q_{bad}^{frozen}, respectively, 0.25 and 0.20), the helix, and the sheet β_{1-5} are formed at lower temperatures in good folders.

V. DISCUSSION

The thermodynamics of the SH3 domain has been widely studied by means of Gō models [26], where only native contacts interact favorably. Conformational samplings of a C_α model where each native contact contributes to the total energy with the same energy $B_0=-1$ show a plot of the specific heat with a single, sharp peak at $0.63|B_0|$ [27]. A modified Gō model where each pair of residues building a native contact interacts with a pair-dependent energy (the average being $B_0=-0.29$ and the standard deviation $\sigma_B=0.37$) displays again essentially a single peak in C_p centered at $0.85|B_0|$ (cf. Ref. [28]). The shape of the specific heat in the present model, where also non-native contacts are considered, is different and much more structured. First, there are a number of peaks which indicate that the Gō interaction oversimplifies the thermodynamics of the chain. Although most of these peaks correspond to transitions that are not found in experiments at biological conditions, this discrepancy makes one suspect that also other features of the thermodynamical states of the model protein can be oversimplified.

Models displaying a simpler geometry (i.e., lattice models) but accounting also for non-native interactions [10,29,30] display a richer thermodynamics than Gō models. As a rule, they show a folding and a coil-globule transition, although it is not always clear how to relate the geometric features of the different phases of a lattice model to that of real proteins. Off-lattice models displaying non-native interactions have been so far applied mainly to simple protein topologies (e.g., helix bundles or β-hairpins) and show simple two-state behavior [31,32]. The model described in this work intends to analyze a more realistic situation where a protein is characterized by a realistic geometry and frustrated interactions.

An interesting feature of the present model is that the biologically relevant unfolded state of the protein (state III of Fig. 3; see also Fig. 4 and Tables II and III) is quite different from a random coil. First, it is rather compact, the average radius of gyration being 12 Å, some 20% larger than the native state. Note that the unfolded state predicted by our model is more compact and much more structured that that given by standard Gō models, which have a $\bar{R}_g \simeq 25$ Å and a total number of contacts that is approximately one-fourth of that in the native state [27]. Second, a number of native and non-native contacts are rather stable in the unfolded state. In particular, the RT loop, the distal loop, and the diverging turn out to be consistently populated.

These results are in agreement with the NMR experiments of α-spectrin SH3 under acidic conditions, which populate a denatured state [33]. This state displaying the nuclear Overhauser effect (NOE) signals in the region of the distal hairpin and of the preceding strand. Moreover, NMR studies of the drkN SH3 domain, an unstable protein which populates the unfolded state under nondenaturing conditions, indicate an even larger abundance of interactions [34] than the α-spectrin experiments, involving the whole regions 9–20 and 25–48. The associated radius of gyration is of the order of 11 Å. The radius of gyration resulting by the implementation of the NOE is $\simeq$11 Å, in agreement with the results of our model. Moreover, 75% of the long-range NOE observed is non-native, a fact that highlights the necessity of accounting for non-native interactions in any model which aims at describing the unfolded state of a protein.

The complicated shape of the specific heat of good folders reveals a hierarchy of energy scales which can be useful to understand the folding of SH3. Some regions of the protein, such as the distal hairpin and the RT loop, are structured even at very high temperature [$q=0.25$ for both structures at $T\approx 0.5$ (state IV); see Table II and Figs. 3 and 4], indicating a remarkable propensity to fold independently of the rest of

the protein. Using the language of [35], we can see these regions as *foldons*.

Following the results obtained with a lattice model and a disordered interaction in Ref. [36], one can interpret these sequence of energy scales from a kinetic point of view, identifying high-temperature states as high-energy conformations at the beginning of the folding dynamics, and low-temperature states as the ending point of the dynamics. From this point of view, the regions of high-temperature conformations displaying native interactions can be regarded as the local elementary structures (LESs) [37] which drive the folding kinetics. Note that this interpretation assigns to nonnative interactions an important role in the folding kinetics, as testified by the fact that their elimination in Gō models affects the whole hierarchy of energy scales. Within the framework of the hierarchical folding mechanism, the RT loop and the distal hairpin act as (closed) LESs in the language of Ref. [38]. Their docking, taking place at the transition between states III and II, gives rise to the (postcritical) folding nucleus (FN), which is the minimum set of native contacts that brings the system over the highest barrier of free energy associated with the folding process.

Our results also show, complementing the findings of our previous work [8], that sequences obtained by minimization of the interaction energy at fixed native conformation not only fold fast but also display realistic thermodynamical features.

Bad sequences, although not being able to fold, display a consistent degree of structure in the same regions of the protein which in good sequences are ordered already at high temperature. This is consistent with the results obtained by means of lattice models [37], where it is seen that contacts within and across local elementary structures are stabilized even in sequences obtained at low evolutionary pressure. In other words, such bad sequences have some of the features typical of good folders, but their energy is not low enough, so that they have to compete with a sea of alternative conformations.

When comparing our results with the random energy model, we find some interesting differences. First of all our model shows a clear folding transition from a nonrandom globular state, which the REM is assumed to require a higher level of side-chain detail [19]. On the other hand we do not see any transition from a random globule to a folded globule for any sequence; sequences either achieve a specific globular configuration, from which they always fold into the native structure, or get trapped into a random globular state. Moreover, the coil-globule transition of our good folders corresponds to a peak in the specific heat, as expected in the case of first-order transitions, in agreement with the nonrandom heteropolymer theory. Our system contradicts also the prediction of the freezing temperature made by the REM (T_{freeze}^{theor} results as much higher than the actual T_{freeze}) and the theoretical equivalence $T_{design}^{cr}=T_{freeze}$, thus questioning the applicability of the self-averaging principle.

Our simulations provide information on the relationship between design temperature (T_{design}) and the thermodynamical behavior of the corresponding sequences in our system.

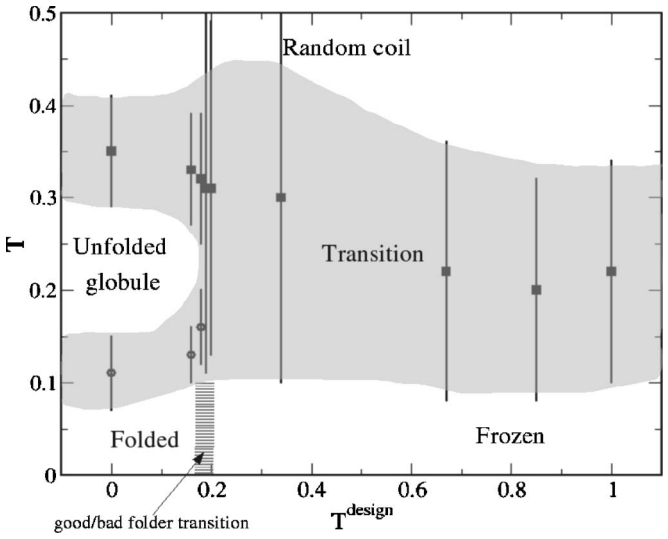

FIG. 11. Phase diagram of the system. The behavior in the conformational space as a function of the design temperature is shown, where the two most relevant transitions are taken into account. The squares identify the centroids of the transitions (top of the C_p peaks for the good folders), while the vertical lines span all over the transition regions for the nine sequences studied. The shaded area marks the expected transition region at any value of the design temperature.

The outcome is summarized in Fig. 11.[4] At high temperatures, sequences designed at any temperature adopt an highly disordered coil configuration. Decreasing the temperature, the behavior becomes sequence dependent. Good sequences first undergo a coil-globule transition to a partially ordered unfolded globule, and from there to the native state, while poorly designed sequences undergo a smooth transition to a frozen set of more or less disordered compact states (R_g of the ground state is $\simeq$10 Å for all the nine sequences). The freezing and folding temperatures remain roughly constant all over the range of T_{design}, while the compaction begins at lower temperatures for higher T_{design}, in such a way that the coil-globule and the globule-frozen transitions "merge" into a wide coil-frozen-globule transition region.

This phase diagram can be compared with the phase diagram arising from the theory of random heteropolymers (see Fig. 1 of Ref. [19]), keeping in mind that the ensembles displayed by our system are not phases in the thermodynamical sense. In overall agreement with the results of the REM, we also observe a random coil, a folded, and a frozen phase. Furthermore, we also find that the boundary between folded and glassy phases is a vertical line in the diagram and the freezing temperature is essentially independent of T_{design}.

Nonetheless, our diagram has important differences from the phase diagram describing the behavior of random heteropolymers. Designed sequences (low T_{design}) display an unfolded globular ensemble showing most of the properties of the unfolded state measured in experiments. Furthermore, the

[4]In that figure we have neglected the first ($V \rightarrow IV$) and that last ($II \rightarrow I$) transictions for the good sequences and focused on the two most relevant ($IV \rightarrow III$ and $III \rightarrow II$).

results shown in Fig. 11 display wide fluctuations between the states, the transition regions occupying most of the diagram. This feature, which is absent in the theoretical diagram calculated in the thermodynamic limit, highlights the important fact that proteins are finite, small systems, and that one should be careful in applying the tools of heteropolymer theory to real proteins.

Another comparison can be made with the mean-field replica calculations performed in Ref. [41] on copolymers, where sequences are not random but weighted according to their E_{targ} by a Boltzmann factor at design temperature T_{design}. The phase diagram that emerges from our simulations matches well with the results of the replica calculations (cf. Fig. 1 of Ref. [41]). In both cases, one observes a folded state, an unfolded globular state, and a frozen state, and the transition between frozen and folded states is essentially temperature independent. The most important difference between the two cases is that replica calculations are a mean-field approximation, and consequently cannot highlight the large fluctuations which characterize the phase diagram of Fig. 11.

VI. CONCLUSIONS

We have used the SH3 domain as a benchmark to test the thermodynamical features of a protein model in which the energy function is nontrivial. Unlike Gō models, this energy function does not contain directly any information on the conformational ground state of the protein, but only through the low (minimized) energy of the sequence in the native conformation. Furthermore, it allows for non-native interactions. The result is a richer set of states than those predicted by Gō models. In particular, the unfolded state of selected sequences is not completely disordered, but is a globule where some of the native contacts are already stabilized. This fact has important implications in the folding kinetics of the protein. The overall picture provided by standard theory of random heteropolymers is verified by our simulations but, again, the model displays richer features, where new equilibrium states are found and where the transition regions play an important role as a consequence of finite-size effects.

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] E. I. Shakhnovich and A. M. Gutin, Protein Eng. **6**, 793 (1993).
[3] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[4] J. Bowie, R. Luthy, and D. Eisenberg, Science **253**, 164 (1991).
[5] J. M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).
[6] C. Micheletti, A. Maritan, and J. Banavar, J. Chem. Phys. **110**, 19 (1999).
[7] C. Micheletti, F. Seno, A. Maritan, and J. Banavar, Proteins **32**, 80 (1998).
[8] A. Amatori, G. Tiana, J. F.-Borg, A. Trovato, L. Sutto, and R. A. Broglia, J. Chem. Phys. **123**, 054904 (2005).
[9] G. Tiana, M. Colombo, D. Provasi, and R. A. Broglia, J. Phys.: Condens. Matter **16**, 2551 (2004).
[10] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).
[11] J. C. Martinez, M. T. Pisabarro, and L. Serrano, Nat. Struct. Biol. **5**, 8 (1998).
[12] J. C. Martinez and L. Serrano, Nat. Struct. Biol. **6**, 11 (1999).
[13] D. S. Riddle, V. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker, Nat. Struct. Biol. **6**, 11 (1999).
[14] W. Guo, S. Lampoudi, and J. E. Shea, Proteins **55**, 395 (2004).
[15] F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, Biophys. J. **83**, 3525 (2002).
[16] I. A. Hubner, K. A. Edmonds, and E. I. Shakhnovich, J. Mol. Biol. **349**, 424 (2005).
[17] R. L. Baldwin and G. D. Rose, TIBS **24**, 26 (1999).
[18] J. Ferkinghoff-Borg, Eur. Phys. J. B **29**, 481 (2002).
[19] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000).
[20] A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994).
[21] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[22] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).
[23] A. Fersht, *Structure and Mechanism in Protein Science* (W. H. Freeman and Company, New York, 1999).
[24] P. Flory, *Statistics of Chain Molecules* (Interscience Publishers, New York, 1969).
[25] T. P. Witelski, A. Y. Grosberg, and T. Tanaka, J. Chem. Phys. **108**, 9144 (1998).
[26] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[27] J. M. Borreguero, N. V. Dokholyan, S. V. Buldryev, E. I. Shakhnovich, and H. E. Stanley, Protein Sci. (to be published).
[28] L. Sutto, R. A. Broglia, and G. Tiana (unpublished).
[29] D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).
[30] C. R. Locker and R. Hernandez, J. Chem. Phys. **120**, 11292 (2004).
[31] A. Irbäck and F. Sjunnesson, Proteins **56**, 110 (2004).
[32] A. Irbäck, B. Samuelsson, F. Sjunnesson, and S. Wallin, Biophys. J. **85**, 1466 (2003).
[33] T. Kortemme, M. J. S. Kelly, L. E. Kay, J. Forman-Key, and L. Serrano, J. Mol. Biol. **297**, 1217 (2000).
[34] Y. Mok, C. M. Kay, L. E. Kay, and J. Froman-Kay, J. Mol. Biol. **289**, 619 (1999).
[35] Anna R. Panchenko, Zaida Luthey-Schulten, and Peter G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **93**, 2008 (1996).
[36] G. Tiana and R. A. Broglia, J. Chem. Phys. **114**, 2503 (2001).
[37] R. A. Broglia and G. Tiana, J. Chem. Phys. **114**, 7267 (2001).
[38] R. A. Broglia, G. Tiana, and D. Provasi, J. Phys.: Condens. Matter **16**, R111 (2004).
[39] S. Knapp *et al.*, Proteins **31**, 309 (1998).
[40] V. V. Filimonov, A. I. Azuaga, A. R. Viguera, L. Serrano, and P. L. Mateo, Biophys. Chem. **77**, 195 (1999).
[41] S. Ramanathan and E. Shakhnovich, Phys. Rev. E **50**, 1303 (1994).