Equilibrium properties of realistic random heteropolymers and their relevance for globular and naturally unfolded proteins

G. Tiana
*Department of Physics, Università degli Studi di Milano and Istituto Nazionale di Fisica Nucleare (INFN),
via Celoria 16, I-20133 Milano, Italy*

L. Sutto
*Spanish National Cancer Research Center (CNIO), Structural Biology and Biocomputing Programme,
Melchor Fernandez Almagro, 3. E-28029 Madrid, Spain*

(Received 5 August 2011; published 12 December 2011)

Random heteropolymers do not display the typical equilibrium properties of globular proteins, but are the starting point to understand the physics of proteins and, in particular, to describe their non-native states. So far, they have been studied with mean-field models in the thermodynamic limit, or with computer simulations of very small chains on lattice. After describing a self-adjusting parallel-tempering technique to sample efficiently the low-energy states of frustrated systems without the need of tuning the system-dependent parameters of the algorithm, we apply it to random heteropolymers moving in continuous space. We show that if the mean interaction between monomers is negative, the usual description through the random-energy model is nearly correct, provided that it is extended to account for noncompact conformations. If the mean interaction is positive, such a simple description breaks out and the system behaves in a way more similar to Ising spin glasses. The former case is a model for the denatured state of globular proteins, the latter of naturally unfolded proteins, whose equilibrium properties thus result as qualitatively different.

DOI: 10.1103/PhysRevE.84.061910 PACS number(s): 87.15.ak, 87.14.E−, 87.15.Cc

I. INTRODUCTION

Random heteropolymers are chains of molecules displaying quenched disordered interactions. Although they usually do not display a dominant equilibrium conformation as globular proteins do, they have been the subject of substantial theoretical interest. Aside from being a model for peptides built out of random sequences of amino acids, like those that are supposed to be involved in prebiotic evolution, random heteropolymers are the starting point to describe the behavior of proteins. Consequently, they have been widely used as a benchmark to study the physics of globular proteins, whose sequence is not random but has undergone natural evolution [1–4], in a way similar to that in which the ideal gas is the starting point to study real gases and the ideal chain is the starting point to study real homopolymers.

The simplest description of a random heteropolymer is through the random-energy model (REM) [5,6], which assumes that the total energy of the system is the sum of a constant, large number n_c of uncorrelated two-body energies, defined by an average ϵ_0 and a standard deviation σ. The model predicts a parabolic entropy

$$S(E) = -\frac{(E - n_c\epsilon_0)^2}{2n_c\sigma^2} \tag{1}$$

down to a ground-state energy $E_c = zN\epsilon_0/2 - N\sigma(z\log\gamma)^{1/2}$, where z is the number of contacts per monomer, so that $n_c = zN/2$, and γ is the number of conformations available to each monomer. The thermodynamics of random heteropolymers has been investigated also in the canonical ensemble with a replica approach [7–9], showing the validity of the REM under the condition that only globular conformations contribute to the partition function [10].

The understanding of the physics of random heteropolymers has been very helpful in the study of globular proteins, that is, of heteropolymers that are nonrandom and that display at biological temperatures a unique equilibrium conformation (the native state). In particular, the $S(E)$ of random heteropolymers has been used to model the unfolded state of proteins. The attractive property of this $S(E)$, and of the associated E_c, which makes this approach useful, is its self-averageness, namely, in the limit of large N it does not depend on the specific realization of the two-body energies, and consequently on the protein sequence. If a protein sequence displays in some conformation an energy lower than the sequence-independent value of E_c, this is the native state of the protein and the system displays the two-state thermodynamics typical of proteins [11,12].

This strategy has been successfully applied to lattice models, designing proteins through the minimization of the energy of the sequence on the wanted native conformation [13], designing potentials to fold specific sequences [14,15], estimating the effect of mutations in the protein sequence [16], or studying natural evolution [17]. In fact, for lattice-model heteropolymers it was shown that the free energy is indeed self-averaging [18] and that the statistical independence of energy levels required by the REM, although not strictly obeyed, is enough to allow protein design [19].

However, lattice models are not very realistic and constrain the polymer much more than what the chemistry of proteins requires. This favors the applicability of the REM, because it depletes the conformational space of correlated conformations. As a matter of fact, the design of the folding sequence in continuous conformational space is much more cumbersome [20,21] and the attempt to obtain a potential to fold protein sequences with variational approaches of the kind of

Refs. [14,15] has been so far frustrated, even using a strong dihedral potential to obtain a latticelike behavior.

The goal of the present work is to investigate the thermodynamics of a heteropolymer model in continuous space, in order to understand whether the REM physics applies and what are the related consequences on the study of proteins. The model is an inextensible chain of beads put at a distance $a = 3.8$ Å interacting through a spherical-well potential

$$U = \sum_{i<j-2}^{N} \left[B(\sigma_i,\sigma_j)\theta(R - |r_i - r_j|) + \frac{1}{\theta(|r_i - r_j| - R_0)} \right], \tag{2}$$

where $B_{\alpha\beta}$ is a 20×20 Gaussian random matrix with mean ϵ and standard deviation $\sigma = 1$, σ_i is a random sequence of integers in the range [1,20] (to mimic natural amino acids), θ is a step function, $R = 5.5$ Å is the range of the interaction, and $R_0 = 4$ Å is the hardcore radius.

The sampling of the conformational space of such a system is not a trivial problem, especially at the low temperatures needed to calculate E_c, and consequently we have first to design a computational strategy to face this problem. This is done in Sec. II. In Secs. III and IV we investigate the applicability of the REM scenario, considering that the typical size N of polypetides and proteins is small (the thermodynamic limit does not make sense in this context) and that here the assumption concerning a constant number of contacts is doubtful. The coil-globule transition is discussed in Sec. V. Particularly interesting is the case of interaction matrices with positive average, which models proteins with low content in hydrophobic amino acids [22]. These proteins are intrinsically disordered and do not display a unique equilibrium state at biological temperature. In Secs. III and VI we show that they populate a phase that is physically distinct than the denatured state of proteins, and try to characterize it.

II. SAMPLING ALGORITHM: ADAPTIVE SIMULATED TEMPERING

The study of the low-energy properties of frustrated systems is computationally challenging [23]. Replica-exchange sampling [24] is a powerful technique, needs very little knowledge of the system to be sampled, but needs large parallel computers to be efficient. On the contrary, simulated tempering [25] is efficient on a single processor, provided that one can tune correctly some parameters of the simulation. In fact, in simulated tempering temperature is regarded as a dynamic variable that changes in a discrete set of values $\{T_i\}$ with rate

$$w(T_i \to T_j) = w_0 \min(1, \exp[-(1/T_j - 1/T_i)E - g_j + g_i]), \tag{3}$$

where w_0 is the rate of attempting a temperature change, E is the energy of the system in the current conformation, and $g(T)$ are dimensionless weights, which are meant to improve the diffusivity of the temperature. The system-dependent parameters to be tuned are then the set of allowed temperatures $\{T_i\}$ and the weights $\{g_i\}$. A uniform sampling of all temperatures is obtained by choosing $g_i = F(T_i)/T_i$, where $F(T_i)$ is the free energy at temperature T_i, which of course is not known in advance.

In order to use simulated tempering in an automatic way, we have developed an adaptive scheme that updates the values of $\{T_i\}$ and of $\{g_i\}$ in a self-consistent way, in the spirit of the approach developed in Ref. [26]. The idea is to carry out a simulated tempering starting at high temperatures, to estimate the density of states of the visited range of energies, and from this to obtain a lower temperature and its weight to continue the tempering efficiently. The procedure is iterated until the system reaches the desired low temperature. Moreover, at each iteration the current set of temperatures and weights is adjusted, exploiting the better knowledge of the density of states obtained as the simulation proceeds.

Specifically, the algorithm works as follows.

(i) A plain Monte Carlo sampling is performed at high temperature T_1 for n_{adj} steps, collecting the histogram of sampled energies.

(ii) A multiple-histogram algorithm [27] extracts the density of states $g(E)$ from the histogram(s) of sampled energy. From $g(E)$, the partition functions and the free energies $F(T)$ are obtained by summation over the sampled energies, and the Boltzmann probability $p_T(E)$ simply from definition. In the iteration following the first one, the histograms used in the multiple histogram are those belonging up to the six previous iterations.

(iii) A new temperature T_i is added below the other(s). The new temperature is chosen in such a way that the mean-field temperature-jump rate $w_{\text{MF}}(T_{i-1} \to T_i)$ is equal to a preset value w_{new}, where

$$\begin{aligned} w_{\text{MF}}&(T_{i-1} \to T_i) \\ &\equiv w_0 \int dE \ \min(1, \exp[(1/T_{i-1} - 1/T_i)E - g_{i-1} + g_i]) \\ &\quad \times p_{T_{i-1}}(E) \end{aligned} \tag{4}$$

and the weights g_i are set to $F(T_i)/T_i$.

(iv) If the number of temperatures n_T used in the simulation is > 3, the temperatures from T_2 to T_{n_T-2} are readjusted, also allowing a decrease of n_T. Operatively, for each possible $n'_T > 3$ the product

$$w_{\text{glob}} \equiv \prod_{i=1}^{n'_T-2} w_{\text{MF}}(T_i \to T_{i+1}) w_{\text{MF}}(T_{i+1} \to T_i) \tag{5}$$

is maximized with respect to the set $[T_2, T_{n'_T-2}]$. The minimum value of n'_T such that w_{glob} is larger than the value it had before is taken as new n_T, together with the associated temperatures and weights.

(v) A simulated tempering with the n_T new temperatures and weights is carried out for a total of $n_T \cdot n_{\text{adj}}$ steps.

(vi) The simulated tempering is considered successful if the fraction of time that the system has spent at each temperature is larger than a preset threshold h_t/n_T and if the jump probability between each pair of consecutive temperatures calculated from the simulation is larger than a preset threshold p_t/n_T. If the simulated tempering fails, the set of temperatures T_i is substituted by the last successful set of temperatures, and the lowest temperature T_{N_T} is raised to half with respect to T_{N_T-1}.

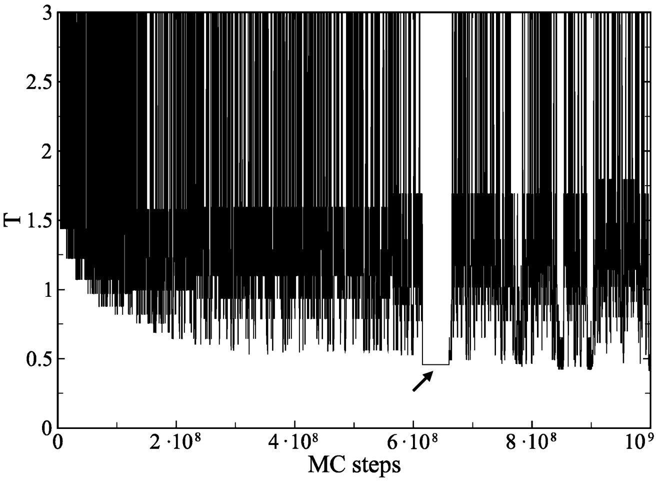

FIG. 1. An example of adaptive simulated tempering. The algorithm adds new temperatures as the simulation goes on. When a set of temperatures fails (see arrow), the system restarts from the last successful set, increasing the lowest temperature.

(vii) If the simulated tempering is considered successful, return to (ii).

Unlike standard simulated tempering, the present algorithm does not rely on the knowledge of thermodynamic features of the specific system, except for the initial temperature T_1. One has only to define some parameters, which control the quality of the simulation, and consequently are weakly dependent on the specific system. The choice we made is $w_{\text{new}} = 4 \cdot 10^{-3}$, $h_t = 0.1$, $p_t = 0.01$, $n_{\text{adj}} = 10^6$, $w_0 = 10^{-4}$. An example of application of the adaptive simulated tempering is given in Fig. 1. For each polymer and each interaction matrix we run the algorithm to obtain an efficient set of temperatures and weights. Then, we repeat thrice a simulated tempering sampling keeping the temperatures fixed, to evaluate the convergence of the thermodynamic quantities.

For each average ϵ of the interaction matrix $(-1,-0.5,0,\text{ and }+1$, while the standard deviation is 1, setting the energy scale of the system) and for each length N of the polymer (20, 25, 30, and 60), we have sampled 20 realizations of the matrix. Almost in all cases we could reach temperatures lower than 0.1. A typical run of the adaptive algorithm for a polymer of 60 residues takes of the order of 10 hours on a single desktop cpu. For chain of length $N = 90$ we were not able to reach full equilibrium at temperatures lower than 0.7, and consequently we discarded such simulations from the analysis.

From these simulations one can obtain the density of states and thus all the other thermodynamic quantities (e.g., the specific heat C_v, as shown in Fig. 2). All C_v obtained for the different realizations of the interaction matrix are quite irregular. All of them show a crowded set of peaks at very low temperatures corresponding to the freezing of the system into the lowest available conformations. These temperatures are $T_g \approx 0.2$ for $\epsilon_0 = 1$, $T_g \approx 0.4$ for $\epsilon_0 = -0.5$, and $T_g \approx 0.7$ for $\epsilon_0 = -1$. In all cases the simulated tempering could visit temperatures below T_g, as expected from the analysis of this algorithm in the case of other frustrated systems [28].

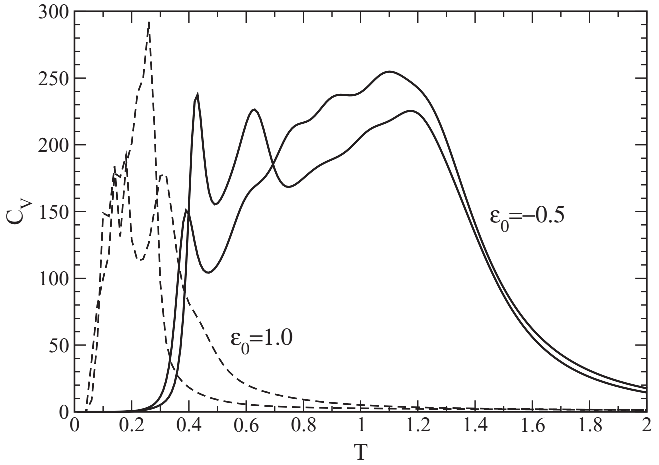

FIG. 2. The specific heat of a polymer interacting with two realizations of the interaction matrix at $\epsilon_0 = -0.5$ and two at $\epsilon_0 = +1$.

III. INTERACTION MATRICES WITH DIFFERENT AVERAGE RESULTS IN TWO DIFFERENT THERMODYNAMIC BEHAVIOURS

The natural order parameter to study the conformational space of random heteropolymers [4,7,9] is the structural overlap

$$q(\alpha,\beta) = \frac{1}{\max(n_c(\alpha),n_c(\beta))} \sum_{i<j} \Delta\left(\left|r_i^\alpha - r_j^\alpha\right|\right)\Delta\left(\left|r_i^\beta - r_j^\beta\right|\right), \tag{6}$$

where $n_c(\alpha) = \sum_{i<j} \Delta(|r_i^\alpha - r_j^\alpha|)$ is the number of contacts of conformation α.

The equilibrium distributions $\overline{p(q)}$ averaged over the interaction matrices are displayed in Fig. 3 for some selected simulations. For negative values of the average ϵ of the

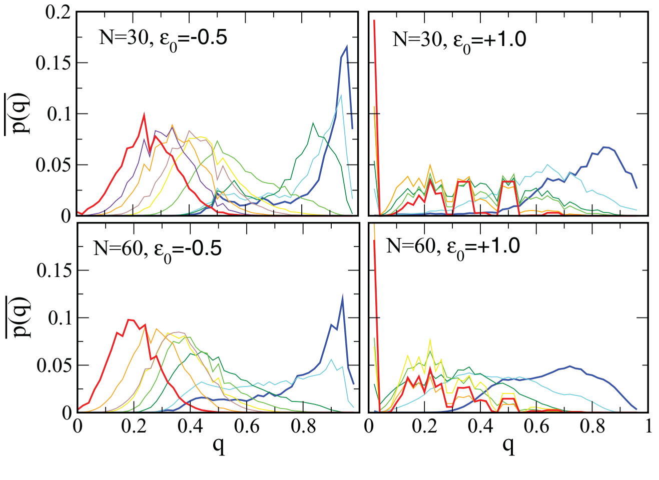

FIG. 3. (Color online) The distribution $\overline{p(q)}$ of the order parameter for two sizes of the polymer ($N = 30$ above and $N = 60$ below) and for two different values of the average ϵ_0 of the interaction matrix ($\epsilon_0 = -0.5$ to the right and $\epsilon_0 = 1$ to the left). In each plot the different distributions are calculated from temperature $T = 0.1$ (blue curve, peaked at $q \approx 1$) to $T = 1.0$ (red curve, peaked at $q \approx 0.2$).

interaction matrix (cf., left panels of the figure) the distribution $\overline{p(q)}$ display two well-defined peaks, one close to $q = 1$ and the other one below $q = 0.5$. As the temperature is decreased, the peak close to $q = 1$ decreases, while the low-q peak increases and moves toward $q \approx 0.2$.

The behavior of the distributions associated with $\epsilon_0 = +1$ is qualitatively different. There is a broad peak at high q, which further broadens increasing N. The top of the peak moves from $q = 0.82$ to $q = 0.74$ as N increases from 30 to 60. The lower part of the distribution does not display a single peak, but a complicated pattern covering the whole range of variability of q. Moreover, there is a sharp peak exactly at $q = 0$. The curves at $\epsilon_0 = -1$ are similar to those at $\epsilon_0 = -0.5$, while those at $\epsilon_0 = 0$ are somewhat in between the two behaviours (data not shown).

The behavior of the distribution at $\epsilon_0 = -0.5$ is typical of frustrated systems undergoing a one-step replica symmetry breaking (RSB) [23], where thermodynamically relevant conformations are either identical to each other (resulting in the $q = 1$ peak) or markedly different (resulting in the peak at $q_0 \ll 1$, cf., Ref. [29]). On the contrary, the bulky shape of $\overline{p(q)}$ and the position of the high-q peak at $q < 1$ observed in the case $\epsilon_0 = +1$ suggest a more complicated pattern of RSB, more similar to the full RSB of Ising spin glasses [23,29] than the one-step RSB observed in our simulations at negative ϵ_0. The sharp peak at $q = 0$ is a polymeric effect: Increasing the temperature, the system stabilizes loose conformations characterized by few contacts (see Sec. V) and the probability that a pair of such conformations share their few contacts is combinatorially low. This effect is of course absent in spin glasses, where the number of contacts is fixed.

A further characterization of the $\overline{p(q)}$ is given by the Binder parameter B, which quantifies the kurtosis of the distributions (see Fig. 4). Although one cannot draw absolute conclusions on the kind of RSB from the shape of $B(T)$ [30], it is still possible to notice that the low-temperature part of $B(T)$ at $\epsilon_0 = 1$ resembles the monotonic shape of the Sherrington-Kirkpatrick model undergoing full RSB (while at higher temperatures departs from that, due to the increasing polymer-swelling effect). On the contrary, at $\epsilon_0 < 0$ it displays bumps similar to those of the three-spin glass undergoing a one-step RSB [31].

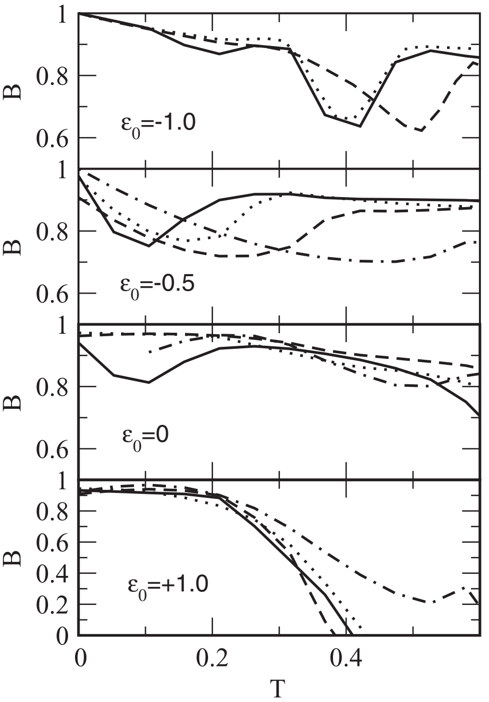

FIG. 4. The Binder parameter B, which is a measure of the kurtosis of the distribution $\overline{p(q)}$, for the polymers controlled by interaction matrices with different averages ϵ_0, as a function of temperature. The curves have been smoothed to facilitate the comparison. The length of the polymer is $N = 20$ (solid curve), $N = 25$ (dotted curve), $N = 30$ (dashed curve), and $N = 60$ (dot-dashed curve).

IV. $\epsilon < 0$ CASE AND ITS RELEVANCE FOR GLOBULAR PROTEINS

The determination of the lowest conformational energy E_c of a polymer controlled by a given interaction matrix is important to understand the folding of globular proteins. In fact, the assumption of a self-averaging behavior of E_c allow to interpret protein evolution as a minimization of their energy E_N in the native conformation and describe the denatured state of the protein with the random-energy model [6,11,32].

To verify these hypotheses, we show in in Fig. 5 the average and the standard deviation of $\epsilon_c \equiv E_c/N$ over the realization of the interaction matrices with varying values of ϵ_0. The average $\overline{\epsilon_c}$ is not independent of N in the range $20 < N < 60$, as required by the random-energy model. The curvature of the curves allows to extrapolate that $\overline{\epsilon_c}$ becomes constant only for $N > 100$. The standard deviation of ϵ_c is a few percent of the average and shows no sign of decreasing with N. This suggests that there is a non-negligible variability of E_c with respect to the interaction matrix and that this variability is not a finite-size effect. For instance, the typical energy scale of interaction between amino acids is kT_{room} and the typical stability of globular proteins (that depends on $E_c - E_N$) is of the order

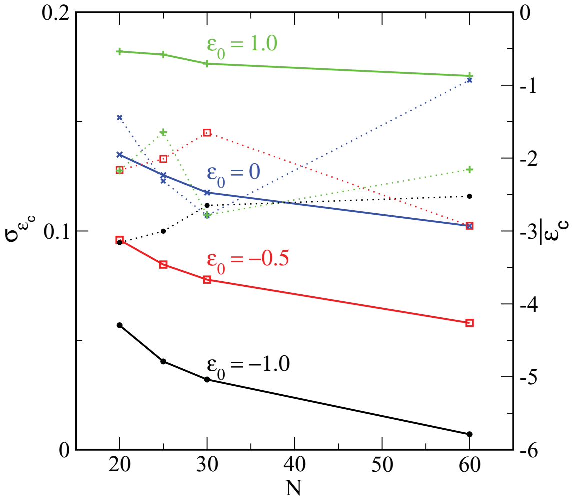

FIG. 5. (Color online) The value of the energy density $\epsilon_c \equiv E_c/N$ of the lowest-energy conformation, averaged over the realizations of the interaction matrices at different ϵ_0 (solid curves, referred to the y axis on the right), and their standard deviation (dotted curves, referred to the y axis on the left).

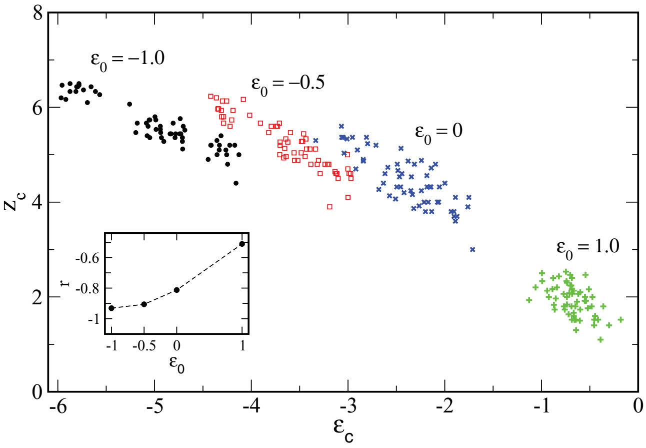

FIG. 6. (Color online) The density of contacts z_c in the lowest-energy conformation, as a function of the associated energy density ϵ_c, for the different realizations of the interaction matrices at various ϵ_0. In the inset, the linear correlation parameter r of the data displayed in the main plot.

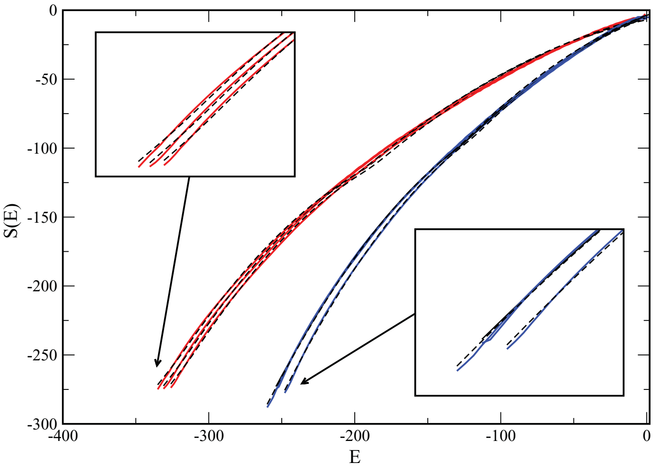

FIG. 7. (Color online) The entropy $S(E)$ as a function of energy for three realizations of random heteropolymers of 60 monomers with average interaction energy $\epsilon_0 = -1$ (red solid curves on the left) and $\epsilon_0 = -0.5$ (blue solid curves on the right). The dashed curves indicate the fit (see text). In the inset, a zoom of the low-energy tails.

of $0.1\,kT_{\text{room}}$ per residue. Consequently, the variability of E_c with respect to the interaction matrix is of the same order of magnitude than the stabilization energy of the protein. Evolution has to do extra work to design proteins with an E_N low enough to be robust against mutations and environmental changes, which could affect the interaction energy. Moreover, becoming larger is not a good evolutionary strategy for proteins to become more stable, as the variability in E_c would also increase.

To investigate the origin of the variability of E_c, we have plotted the density of contacts $z = 2n_C/N$ within the lowest-energy conformation with respect to its energy density ϵ_c (see Fig. 6). First of all, one can notice that even for $\epsilon_0 < 0$ the density of contacts varies within a 30% from polymer to polymer, indicating that the ground-state conformation is not completely compact. In this case, the value of ϵ_c is well correlated to z (cf., the correlation coefficient in the inset of Fig. 6), suggesting that the variability in E_c is due to the different number of contacts that low-energy conformations can accommodate.

The entropy function whose endpoint is E_c is given in Fig. 7 for a few representative cases. The curve $S(E)$ is variable with respect to the different realizations of the interaction matrix. Differently than the predictions of the REM, $S(E)$ cannot be fitted by the parabola of Eq. (1). As a matter of fact, the REM assumes a constant number of contacts, while polymers display a nontrivial distribution of the number of contacts, which is also responsible for the coil-globule transition [33].

The shape of $S(E)$ seems to display two different behaviors at low and high energies, which can be captured by a two-REM description, that is

$$S(E) = \ln\left[\alpha \exp\left(-\frac{(E - zN\epsilon_0/2)^2}{zN\sigma^2}\right) + \alpha' \exp\left(-\frac{(E - z'N\epsilon_0'/2)^2}{z'N\sigma'^2}\right)\right], \qquad (7)$$

where N is the length of the chain, $\epsilon_0 = -1, -0.5$, and $\sigma = 1$ are the mean and the standard deviations of the interaction matrices. The fit over the other parameters gives a reduced χ^2, which ranges between 0.98 and 3.07 for $N = 60$ and worsen for smaller chains. This suggests that it is possible to describe effectively the states of the heteropolymer as a superposition of two set of conformations building out, respectively, z and z' interactions per monomer, and corresponding to the two Gaussians in Eq. (7). The fit gives values of z and z' whose means over the realizations is $\overline{z} = 3.22 \pm 0.12$ for $\epsilon_0 = -0.5$ and $\overline{z} = 4.10 \pm 0.14$ for $\epsilon_0 = -1$. This indicates compact conformations, but not fully compact (because $\overline{z}$ increases with ϵ_0). Moreover, the variability over the different realizations of the interaction matrix is very small. On the other hand, to the high-energy part of $S(E)$ is associated a $\overline{z'} = 1.52 \pm 0.35$ and $\overline{z'} = 0.34 \pm 0.22$ for $\epsilon_0 = -0.5$ and $\epsilon_0 = -1$, respectively. These correspond to more swollen conformations and suffer a much larger variability among realizations of the interaction matrix. To fit properly the high-energy part of the computed $S(E)$ it is also necessary to use values of ϵ_0' and σ' different from the values set for the interaction matrix (the fit with the actual values $\epsilon_0' = -1, -0.5$, and $\sigma' = 1$ give a reduced $\chi^2 > 30$).

The picture that emerges is that for negative values of ϵ the REM fails because the underlying hypothesis of a constant number of contacts does not apply. Nonetheless, the system can be effectively described as built out of two sets of conformations, different for the (constant) number of contacts, each of them displaying a REM-like behavior. The parameters controlling the more compact conformations are rather independent on the interaction matrix, while those controlling the swollen conformations are not.

The low-energy end of the $S(E)$ curve is irregular and departs slightly from the REM behavior in a matrix-dependent way (see insets of Fig. 7). This irregularity involves only an interval of few σ and sets the actual value of E_c below that predicted by the REM. Thus, the major determinant in the

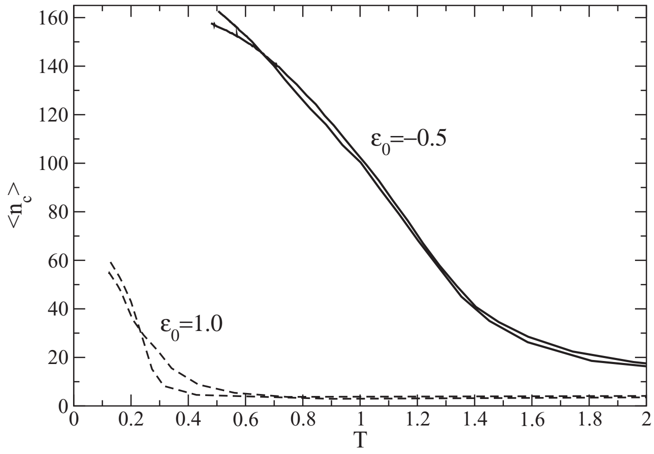

FIG. 8. The average number of contacts with respect to temperature.

variability of E_c seems to be the shape of the high-energy part of $S(E)$.

V. COIL-GLOBULE TRANSITION

The heteropolymers interacting with $\epsilon_0 < 0$ display a broad peak in the specific heat at high temperature (see Fig. 2), which corresponds to the midpoint in the decrease of the number of contacts in the chain (see Fig. 8), and consequently to a transition from globular to coil states. In the proteinlike range of chain lengths the width of the peak in C_v is quite constant with respect to the realization of the interaction matrix (e.g., at $\epsilon_0 = -0.5$ $\Delta T = 0.78 \pm 0.06$ at half height for $N = 60$ and $\Delta T = 0.82 \pm 0.14$ for $N = 30$) but does not display the $N^{-1/2}$ behavior associated to the coil-globule transitions of homopolymers.

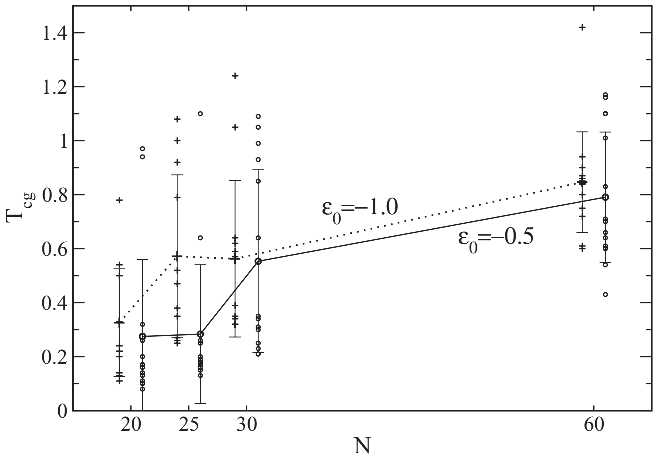

FIG. 9. The coil-globule transition temperatures for the different realizations of the interaction matrix with $\epsilon_0 = -1$ (crosses) and $\epsilon_0 = -0.5$ (circles). The dotted and solid lines indicate the average over realizations, while the error bars the associated standard deviation. The set of points at different values of ϵ_0 and a given value of N are slightly displaced along the horizontal axis to allow a clearer identification.

On the other hand, the transition temperature (operatively defined as the temperature T_{cg} corresponding to the top of the highest-T peak in C_v) is rather realization dependent, and does not show any clear trend in decreasing its variability with respect to the length of the heteropolymer, as shown in Fig. 9.

The replica approach of Ref. [34] highlights a coil-globule transition only for $\epsilon_0 > 0$, due to the strong hypotheses on the density of the chain. A Flory-Huggins description of a homopolymer corrected with the effective second virial coefficient introduced in Ref. [34] predicts a coil-globule transition, which depends on the number of contacts of the most compact conformation [21]. The values of T_{cg} calculated according to Ref. [21] from the number of contacts displayed in Fig. 6 correlate poorly ($r = 0.32$) with the values of T_{cg} of Fig. 9. This suggests that simple mean-field theories are not able to capture the large variability of the coil-globule transition temperatures.

A high-temperature-expansion approach identifies a first-order transition between a frozen globule and a random coil, and a second-order transition between a random globule and a coil [35]. Due to the limited size of the heteropolymers that can be treated computationally, we are not able to investigate the order of the transition.

A striking feature that emerges from these calculations is that the coil-globule transition is very broad in the range of polymer lengths corresponding to single-domain proteins. This means that it is very likely that the denatured state of proteins belong to the transition region, which also depends on the details of the interaction between amino acids. As a consequence, one expects a high variability in the size of the denatured state of proteins.

VI. $\epsilon > 0$ CASE AND ITS RELEVANCE FOR NATIVELY UNFOLDED PROTEINS

As discussed in Sec. III, in the case of positive mean of the interaction matrix the REM scenario does not hold. In fact, the shape of $S(E)$ associated with the different realizations of the matrix with mean $\epsilon_0 = 1$ displays an irregular behavior (see Fig. 10). These curves cannot be fitted by a parabola or

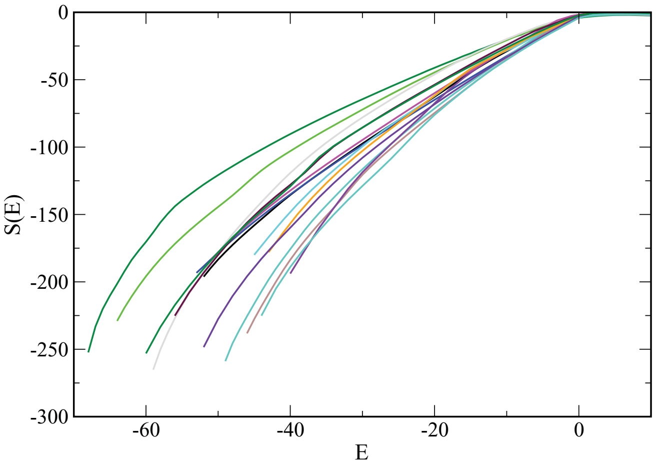

FIG. 10. (Color online) The entropy $S(E)$ as a function of energy for fifteen realizations of random heteropolymers of 60 monomers with average interaction energy $\epsilon_0 = 1$.

by Eq. (7), the reduced χ^2 being larger than 30, in agreement with a complicated RSB pattern.

The ground-state conformations display a spread that is comparable with that of negative ϵ_0 (cf., Fig. 5), something that is somewhat unexpected, due to the irregularity of the associated $S(E)$. Such conformations are still globular [with a z of the order of 2 (see Fig. 6)] even if much less compact than those with $\epsilon_0 < 0$. But differently from that case, now the number of contacts decreases drastically above E_c. From a canonical-ensemble point of view, the specific heat displays a single broadened (and quite irregular) peak at low temperatures (see Fig. 2), which inevitably marks the glassy transition. But in the same range of temperatures ($T = 0.1$–0.4) the average number of contacts decreases to values typical of coils, consequently the system jumps from a glassy globule to a random coil.

The study of random heteropolymers interacting through a matrix with positive mean is interesting because they represent the lowest-order approximation of natively unfolded proteins. This class of proteins does not display a unique native conformation in solution, but displays biological activity either when unstructured, or getting structured (or partially structured) upon binding other molecules [36]. The low content of hydrophobic amino acids in these kinds of proteins [22] suggests that the average interaction is much less attractive than that of globular proteins. The hydrodynamics radius of natively unfolded proteins in solution is either that of a random coil or that of a molten globule, depending on the specific protein [36]. This indicates that biological temperature is likely to lie in or close to the transition region ($T = 0.2$–0.4 in Fig. 8), also in the case of coil proteins, which get promptly compacted upon binding.

A consequence of this scenario is that the free-energy profile of natively unfolded proteins is quite different from that of unfolded globular proteins. As in Ising spin glasses, where a full RSB transition applies, the free energy displays a hierarchical tree of states at all energy scales, giving rise to conformational substates reminiscent of those observed in myoglobin but at very low temperature [37]. In other words, while the denatured state of globular proteins is expected to display a set of conformations, which can only be completely different from each other (i.e., are different at the length scale of the whole protein), natively unfolded proteins are expected to populate conformations that are different from each other at all possible length scales.

As a matter of fact, several experiments carried out on alpha-synuclein, a natively unfolded protein, give results that are consistent with this picture. Single-molecule Förster resonance energy transfer experiments at room temperature provide broadened distributions of distances between pairs of residues [38]. While the authors of this work comment that (they) cannot eliminate the possibility that peak broadening results from alpha-synuclein sampling of two or more specific conformations with close mean energy-transfer-efficiency values that are not resolved in their measurement, a full RSB scenario would explain well such a peak broadening. A study of the same protein with fluorescent energy transfer combined with electron transfer measurement also shows distance distributions that are significantly broadened and appear to be continuous at the fitting resolution of 2 Å [39], compatibly with the scenario we suggest.

Moreover, nuclear magnetic resonance experiments on another natively unfolded protein, HIV-1 Tat, show many weak cross peaks, more abundant and more broadened than expected from a 92-residue protein [40]. Again, this is compatible with an energy landscape whose roughness involves all length scales.

The full-RSB scenario has deep consequences in the way one models the experimental data in order to obtain structural information on a natively unfolded protein. In fact, one approach is to try to obtain well-defined clusters of conformations, which are overall compatible with the data [41]. If a one-step RSB applied, such clusters would be naturally defined as different on the length scale of the whole protein. But in a full-RSB scenario, where differences apply over a continuum of length scales, one expects the classification of probable conformation to be more complicated, due to the absence of a natural length scale to distinguish between them.

Another interesting phenomenon that takes place in frustrated systems is the existence of slow and multiple relaxation time scales and aging [29]. It was shown that these phenomena affect both RSB schemes, although at different scales [42]. Anyway, they have been observed in the case of natively unfolded proteins on the time scale of milliseconds to microseconds [40,43], and not in the denatured state of globular proteins up to the time scale of nanoseconds [44], suggesting a marked quantitative difference between the two cases.

VII. CONCLUSION

The study of random heteropolymers with continuous degrees of freedom and with size comparable to that of small globular proteins is computationally demanding, but important to complement the available mean-field theories and the simulations done with lattice models and very short polymers. If the interaction energy between monomers is negative, random heteropolymers display a one-step replica symmetry breaking at low temperatures, and the density of states can be described by a modified random-energy model, which accounts also for noncompact conformations. The effect of these swollen conformations is to cause a variability in the lowest-energy E_c available to the system with respect to the details of the interaction potential. Such a variability is not negligible and should be accounted for in the *de novo* design of globular proteins and in the study of the effect of mutations in protein sequences.

On the other hand, if the mean interaction energy is positive, a more complicated replica symmetry breaking pattern takes place at low temperatures, suggesting that the free-energy profile is now more complicated. This scenario is expected to apply to natively unfolded proteins, whose residues are on average less hydrophobic than those of globular proteins.

We think that the self-adjusting simulated-tempering algorithm employed for this investigation can be useful to study the low-energy properties of other systems without the need of massive parallel computers.

[1] E. I. Shakhnovich and A. M. Gutin, J. Phys. A **22**, 1647 (1989).

[2] J. D. Bryngelson and P. G. Wolynes, Biopolymers **30**, 177 (1990).

[3] S. Ramanathan and E. I. Shakhnovich, Phys. Rev. E **50**, 1303 (1994).

[4] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Phys. Rev. E **51**, 3381 (1995).

[5] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[6] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).

[7] A. M. Gutin and E. I. Shakhnovich, Sov. Phys. JETP **69**, 1185 (1989).

[8] S. Franz, M. Mezard, and G. Parisi, Int. J. Neural Sys. **3**, 195 (1992).

[9] C. D. Sfatos, A. M. Gutin, and E. I. Shakhnovich, Phys. Rev. E **48**, 465 (1993).

[10] J. Wilder and E. I. Shakhnovich, Phys. Rev. E **62**, 7100 (2000).

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

[12] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992).

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

[14] F. Seno, C. Micheletti, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **81**, 2172 (1998).

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

[16] G. Tiana, R. A. Broglia, H. E. Roman, E. Vigezzi, and E. I. Shakhnovich, J. Chem. Phys. **108**, 757 (1998).

[17] G. Tiana, R. A. Broglia, and E. I. Shakhnovich, Proteins **39**, 244 (2000).

[18] J. Chuang, A. Yu. Grosberg, and M. Kardar, Phys. Rev. Lett. **87**, 078104 (2001).

[19] V. S. Pande, A. Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).

[20] A. Amatori, G. Tiana, L. Sutto, J. Ferkinghoff-Borg, A. Trovato, and R. A. Broglia, J. Chem. Phys. **123**, 054904 (2005).

[21] A. Amatori, J. Ferkinghoff-Borg, G. Tiana, and R. A. Broglia, Phys. Rev. E **73**, 061905 (2006).

[22] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, Proteins **42**, 38 (2001).

[23] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glasses and Beyond* (World Scientific, Singapore, 1987).

[24] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).

[25] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).

[26] J. Ferkinghoff-Borg, Eur. Phys. J. **29**, 481 (2002).

[27] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[28] E. Marinari, G. Parisi, and J. J. Ruiz-Lorenzo, Phys. Rev. B **58**, 14852 (1998).

[29] K. H. Fischer and J. Hertz, *Spin Glasses* (Cambridge University Press, Cambridge, 1991).

[30] K. Hukushima and H. Kawamura, Phys. Rev. E **62**, 3360 (2000).

[31] M. Picco, F. Ritort, and M. Sales, Eur. Phys. J. B **19**, 565 (2001).

[32] E. I. Shakhnovich and A. M. Gutin, Protein Engin. **6**, 793 (1993).

[33] M. P. Taylor, W. Paul, and K. Binder, Phys. Procedia **4**, 151 (2010).

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

[35] V. S. Pande, A. Yu Grosberg, and T. Tanaka, J. Chem. Phys. **107**, 5118 (1997).

[36] V. N. Uversky, Protein Science **11**, 739 (2002).

[37] H. Frauenfelder, F. Parak, and R. D. Young, Annu. Rev. Biophys. Biophys. Chem. **17**, 451 (1988).

[38] A. J. Trexler and E. Rhoades, Biochem. **48**, 2304 (2009).

[39] J. C. Lee, H. B. Gray, and J. R. Winkler, J. Am. Chem. Soc. **127**, 16388 (2005).

[40] S. Shojania and J. D. O'Neal, J. Biol. Chem. **281**, 8347 (2006).

[41] T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, M. Tyers, and J. D. Forman-Key, Structure **18**, 494 (2010).

[42] G. Ben Arous, A. Bovier, and J. Černy, Commun. Math. Phys. **282**, 663 (2008).

[43] K. P. Wu, S. Kim, D. A. Fela, and J. Baum, J. Mol. Biol. **378**, 1104 (2008).

[44] D. Nettels, A. Hoffmann, and B. Schuler, J. Phys. Chem. B **112**, 6137 (2008).