Role of bulk and of interface contacts in the behavior of lattice model dimeric proteins

G. Tiana,[1,2] D. Provasi,[1,2] and R. A. Broglia[1,2,3]
[1]*Department of Physics, University of Milano, via Celoria 16, 20133 Milano, Italy*
[2]*INFN, Sezione di Milano, Milano, Italy*
[3]*The Niels Bohr Institute, University of Copenhagen, Bledgamsvej 17, 2100 Copenhagen, Denmark*
(Received 6 November 2002; published 14 May 2003)

Some dimeric proteins first fold and then dimerize (three-state dimers) while others first dimerize and then fold (two-state dimers). Within the framework of a minimal lattice model, we can distinguish between sequences following one or the other mechanism on the basis of the distribution of the ground state energy between bulk and interface contacts. The topology of contacts is very different for the bulk than for the interface: while the bulk displays a rich network of interactions, the dimer interface is built up of a set of essentially independent contacts. Consequently, the two sets of interactions play very different roles both, in the folding and in the evolutionary history of the protein. Three-state dimers, where a large fraction of energy is concentrated in few contacts buried in the bulk, and where the relative contact energy of interface contacts is considerably smaller than that associated with bulk contacts, fold according to a hierarchical pathway controlled by local elementary structures, as also happens in the folding of single-domain monomeric proteins. On the other hand, two-state dimers display a relative contact energy of interface contacts, which is larger than the corresponding quantity associated with the bulk. In this case, the assembly of the interface stabilizes the system and leads the two chains to fold. The specific properties of three-state dimers acquired through evolution are expected to be more robust than those of two-state dimers; a fact that has consequences on proteins connected with viral diseases.

DOI: 10.1103/PhysRevE.67.051909 PACS number(s): 87.15.Cc

I. INTRODUCTION

Dimers are a rather common structure adopted by proteins to perform their biological activity. They are proteins whose native conformation is a globule build out of two disjoint chains. In particular, homodimers are dimers composed of two identical sequences.

The importance of the study of the evolutionary properties of homodimers is connected with the fact that enzymes of this type produced by viral agents (e.g., HIV-1-protease) are able to display very fast evolutionary patterns to escape from the pressure exerted by the immune system and by drugs. The knowledge of the evolutionary properties of dimers can thus be of help in designing a strategy to deal with the associated diseases.

Some of the known homodimers, such as *E. Coli* Trp repressor [1], fold according to a three-state mechanism, where first the denaturated chains of the monomers assume conformations rich of native structures independently of each other, and subsequently the two parts come together to form the dimer. A different behavior is displayed by, for example, P22 Arc Repressor, whose chains dimerize without populating any monomeric nativelike intermediate (two-state process). In this case one can only detect the unfolded monomers and the native dimers [2]. A class by itself is composed by dimers that have evolved by swapping entire domains of each monomer (''domain swapping dimers,'' Ref. [3]) so that a segment of a monomer is replaced by the same segment of the other monomer.

Making use of a simple model of protein folding, it has been previously shown [4] that the folding of dimeric proteins is, to a large extent, controlled by their ground state energy, as already found for monomeric proteins. In the case of monomers, good folders are those sequences that display a low energy in the native state (i.e., an energy below a threshold E_c, quantity determined mainly by the length of the chain and by the statistical properties of the interaction matrix [5,6]). This is also found to be true for dimeric sequences. Moreover, in this case, one can distinguish between two- and three-state dimers by studying how the relative contact energy [7] is distributed between the bulk (i.e., the interactions between residues belonging to the same monomer divided by the corresponding number of contacts) and the interface (i.e., the interactions between residues belonging to different monomers divided by the number of interface contacts).

From the evolutionary point of view, it is possible to obtain one or the other behavior concentrating the evolutionary pressure either on the bulk or on the interface amino acids. If the evolutionary pressure, and consequently the stabilization energy per contact, is concentrated on the bulk, the dimer displays a three-state folding mechanism. The resulting behavior is, in this case, similar to that displayed by monomeric chains, where local elementary structures (LES), formed and stabilized at early stages of the folding process, essentially control the folding process. The assembly of the LES into the folding nucleus allows the chain to overcome the main free energy barrier in the path towards the monomeric native state [6]. For three-state dimers there is an additional step in the folding hierarchy, corresponding to the assembly of the twofolded monomers into the native dimer. On the other hand, if the overall relative contact energy is low but the relative energetic balance favors the interface contacts, a two-state dimerization process takes place, where first the interface is built and then the other residues fold around it. In this case a larger number of residues, than in the

case of three-state dimers, are involved in the folding process; essentially all these belong to the interface, rather than just few, highly conserved, strongly interacting, residues as in the case of three-state folders.

The two cases, however, are not symmetric. The bulk of the two dimers contains a rich network of interactions, while the interface is kept together by bonds that are, to a large extent, independent on each other. This topological difference gives rise to different evolutionary properties for the residues at the interface with respect to those buried inside the monomers. The evolutionary features of model dimers can be compared with those of real dimers through the analysis of the conservation patterns in families of analogous proteins.

The model we use to study homodimers is largely employed in the literature [5,8,9] because, in spite of the simplifications introduced in the description of the protein so as to make the calculations feasible, it still contains the two main ingredients which form the base of the distinctive properties of proteins: polymeric structure and disordered interactions among the amino acids [10]. According to this model, a protein is a chain of beads on a cubic lattice, each bead representing an amino acid that interacts with its nearest neighbors through a contact energy as described in Ref. [11] (for details about the model see, e.g., Refs. [5,12]).

The lattice model has been shown to be qualitatively in agreement with the experimental data in describing a large number of thermodynamical, kinetical, and evolutionary features of monomeric proteins (see, e.g., the reviews [13,14]), while it neglects the structural details of the protein chain. On the other hand, a more realistic model describing at full-atomic or partially full-atomic level the protein would be too demanding computationally in order to collect enough statistics of the folding process of a dimer [15]. Anyway, the purpose of the present work is to extend the results found with minimal models from monomeric to dimeric proteins and to gain insight into the general mechanism of folding, rather than to describe the folding of some specific protein.

II. SELECTING FOLDING SEQUENCES BY REPRODUCING EVOLUTION

In the case of monomeric sequences, the canonical ensemble associated with the space of sequences for a fixed native conformation has been proved useful [19] in selecting proteins with given thermodynamical and kinetical properties, due to the fact that these properties are essentially determined by the native (ground state) energy. In this context, the energy of a sequence is controlled by an ''evolutionary'' temperature τ, in such a way that the probability of selecting a sequence with energy E is proportional to $\exp(-E/\tau)$. τ is an intensive variable that plays the role of temperature and gives the degree of bias towards low-energy sequences. In the evolutionary context, τ has the meaning of selective pressure with respect to the ability of the protein to fold [20]. In particular, for values of τ lower than temperature $\tau^c \equiv (\partial S/\partial E|_{E=E_c})^{-1}$, the average energy of the selected sequences is lower than E_c, which is the energy that separates sequences with a unique and stable native conformation and is able to find it rapidly from random heteropolymers, which undergo a lengthy and nonunique compaction process [6,19].

In the case of homodimers, we want not only to distinguish good folders from bad folders, but also want to separate those displaying a two-state folding mechanism from those displaying a three-state folding mechanism. This is done by controlling separately the energy E_1 associated with the interaction of amino acids belonging to the same monomer and the energy E_t associated with the interaction of amino acids belonging to different monomers. Accordingly, we select sequences $\{s\}$ with probability

$$p(\{s\}) = \frac{1}{Z}\exp\left(-\frac{H_1}{\tau_1}-\frac{H_t}{\tau_t}\right), \tag{1}$$

where H_1 and H_t are, respectively, the bulk and the interface energy of the sequence on the chosen native conformation and $Z=\Sigma_{\{s\}}\exp(-H_1/\tau_1-H_t/\tau_t)$ is the normalization constant. Parameters τ_1 and τ_t are intensive variables that give the degree of evolutionary pressure on the bulk and on the interface.

The nonequilibrium distribution $p(\{s\})$ of Eq. (1) is the stationary distribution that maximizes the Shannon entropy of the system at the given values of the average bulk and interface energies. In other words, we are interested in a distribution that does not depend on time and that does not give all microscopic details of the system (i.e., minimizing the information, that is maximizing the entropy), except for the two average energies over which we want to have control. Of course, if the two average energies are set to different values, this is a nonequilibrium distribution, implying that each of the two parts of the system is in contact with its own thermal bath.

Anyway, it is possible to describe the system in a way that is formally similar to that of the equilibrium canonical ensemble. If we call p a generic distribution of states of the system, we can define the average energy functional $E_1[p] = \Sigma H_1 p$ and $E_t[p]=\Sigma H_t p$ of the two parts of the system and the entropy functional $S[p]=-\Sigma p \ln p$, which indicates the information we have about the system.

Among all possible distributions p, we are interested in the stationary distribution p^* that minimizes the information (maximizing S) at fixed values of the average energy $E_1[p]$ and $E_t[p]$; values that we call E_1^* and E_t^*, respectively. In keeping also with the constraint $\Sigma p=1$, one minimizes the functional

$$S[p]-\alpha(E_1[p]-E_1^*)-\beta(E_t[p]-E_t^*) - (\ln Z+1)\left(\sum p-1\right), \tag{2}$$

where α, β, and λ are the Lagrange multipliers that fix the value of the average energies and of the normalization factor, respectively. Setting the derivative of this functional to zero gives the required expression for the stationary probability, that is,

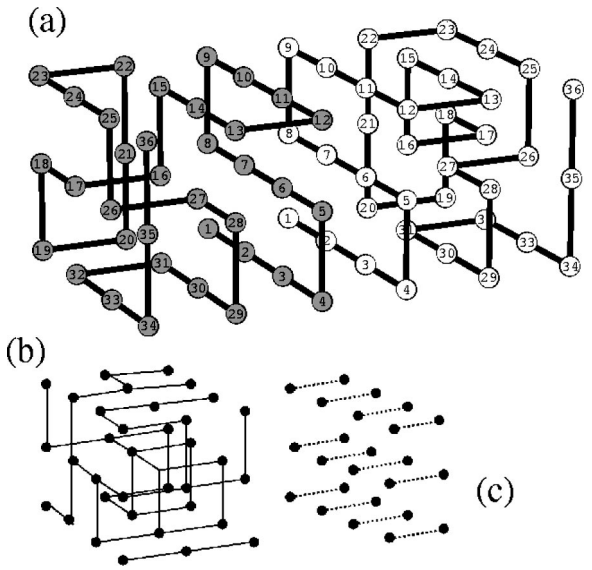

FIG. 1. (a) The native conformation used in the present calculations. (b) The rich network of 40 contacts formed by the residues belonging to one of the monomers displayed in (a). (c) The 12 contacts between the two monomers.

$$p^* = \frac{1}{Z}\exp(-\alpha H_1 - \beta H_t), \tag{3}$$

where the evaluation of the Lagrange multipliers gives $Z = \Sigma \exp(-\alpha H_1 - \beta H_t)$, which has the form of a partition function, and

$$\left.\frac{\partial S}{\partial E_1}\right|_{E_1^*} = \alpha; \quad \left.\frac{\partial S}{\partial E_t}\right|_{E_t^*} = \beta. \tag{4}$$

In parallel with equilibrium thermodynamics, we call ''temperature'' the inverse of the two Lagrange multipliers, $\tau_1 = \alpha^{-1}$ and $\tau_t = \beta^{-1}$. The simultaneous action of the two evolutionary temperatures induces, in the design of dimers, a selective bias towards sequences displaying a conspicuous low energy in the native conformation, in a similar way that a single evolutionary temperature does in the case of the design of single-domain monomeric proteins [6]. Lowering τ_1 or τ_t increases the pressure set in the bulk or on the interface, respectively.

To implement this procedure, we use a multicanonical technique [21,22] in the space of sequences, for a fixed dimeric native conformation. First, we select a target conformation built of two identical parts (in the present case each being a 36-mer) having a face in contact, in such a way that the overall structure is symmetrical with respect to the interface [23] as, e.g., in Fig. 1(a), choosing a realistic ratio among the different kinds of amino acids [24]. We swap amino acids, thus keeping the ''wild-type'' concentration fixed; accepting or rejecting the swap with the help of a multicanonical algorithm. In this way one can select a set (composed typically of 10^4 elements) of evolutionary uncorrelated sequences $\{s\}$ corresponding to a given pair of values τ_1 and τ_t, all designed under similar evolutionary conditions.

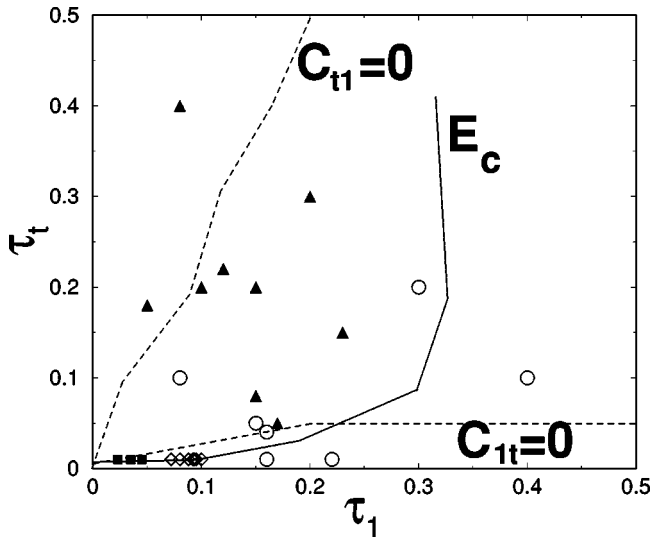

FIG. 2. Dynamical phase diagram of dimeric sequences selected at different values of the evolutionary temperatures τ_1 and τ_t. Solid squares indicate two-state folding sequences; solid triangles indicate three-state folding sequences; empty symbols label sequences that display specific (diamonds) and unspecific (circles) aggregation. The solid curve corresponds to the loci of the set of values of (τ_1, τ_t) associated with the total energy E_c, while dashed curves limit the area outside which the specific heat associated with either the volume or the interface degrees of freedom is negative.

III. PROPERTIES OF THE SPACE OF SEQUENCES

The dynamical simulation of the folding process of sequences selected at different values of τ_1 and τ_t produces the phase diagram displayed in Fig. 2, testifying to the fact that the folding properties of a sequence depend on its ground state energy. One can identify four areas in the phase diagram, corresponding, respectively, to a two- or to a three-state folding behavior and to two different kinds of aggregations: specific (i.e., depending on certain contacts) or not (i.e., a random collapse).

The kinetic properties of the selected sequence have been analyzed by means of Monte Carlo (MC) dynamical simulations using the standard move set (composed of head or tail, corner flip, and crankshaft moves), which has been shown to reproduce well the dynamics of a polymer chain (cf. Ref. [25] and Appendix B of Ref. [4]). For each sequence, 20 simulations of 60×10^6 MC steps have been performed at constant temperature. The procedure has been repeated at three different temperatures ($T = 0.24$, $T = 0.28$, and $T = 0.32$, in the units of the interaction matrix [11], in the neighborhood of the folding temperature, cf. Ref. [4]). The two- and three-state characters of a given sequence have been studied by inspecting the similarity of the bulk and of the interface to the native conformation when the interface and the bulk, respectively, are formed. For more details about the folding procedures see Ref. [4].

Sequences selected at very low [26] values of τ_1 and τ_t fold according to a two-state mechanism; first dimerizing and then folding to their dimeric native state. Sequences following this behavior are indicated with solid squares in Fig. 2. Rising τ_t produces sequences that fold according to a three-state mechanism; first folding to the monomeric native state and then dimerizing (solid triangles). Leaving fixed τ_t, we find that the upper limit of τ_1 still leading to designed sequences folding according to any of these two paradigms corresponds to that value for which total energy $2E_1+E_t=E_c$ (≈ -35. Throughout this paper, energy is given in units of the interaction matrix [11], i.e., $RT_{room}=0.6$ kcal/mol). Examples of sequences folding according to the two mechanisms are

$$S_1 \equiv \text{VLNLGNFVGGHCRYDMEASLWTAKPKPTIRISEADQ (two-state),}$$

$$S_3 \equiv \text{NTKPVERNCTRVIDGDFALYSGAGSMKLQEHLWPIA (three-state),}$$

whose energies are $E_{designed}$ $(=2E_1+E_t)=-36.1$ $(E_1=-15.50, E_t=-5.11)$ and $E_{designed}=-36.8$ $(E_1=-16.47, E_t=-3.89)$, respectively.

Keeping τ_t low and increasing τ_1 (so as to meet the curve associated with E_c) leads to sequences that aggregate (empty diamonds in Fig. 2). An example of such sequences is provided by

$$S_5 \equiv \text{CGNLVNGHVFGLASMKPRPSDIQWTREAIODYELTA (aggregation),}$$

with $E_{designed}=-35.0$ $(E_1=-14.95, E_t=-5.11)$.

Aggregation takes place because the interface becomes too reactive with respect to the bulk. For evolutionary temperatures outside the area defined by the ordinate $(\tau_1=0)$ and E_c (solid curve in Fig. 2), the equilibrium state is, at any temperature, either a disordered clump made up of the two chains, or a state where the two chains are separated and unfolded, depending on their concentration (cf. Sec. V).

The central role played by the relative contact energy associated with the bulk and with the interface can be assessed from Fig. 3. Due to the fact that typical homodimers display more bulk contacts than interface contacts [e.g., in the case under discussion these are 80 and 12, respectively, cf. Figs. 1(b) and 1(c), respectively], it is not very useful to compare the total values $2E_1$ and E_t, but rather the average values $\epsilon_1=2E_1/$(No. of bulk contacts), $\epsilon_t=E_t/$(No. of interface contacts), $\epsilon_{designed}=E_{designed}/$(total No. of contacts), and $\epsilon_c=E_c/$(total No. of contacts).

Figure 3 suggests that provided $\epsilon_{designed}<\epsilon_c$, the system folds. It folds as a three-state dimer if $\epsilon_1<\epsilon_t$, and as a two-state dimer if $\epsilon_1>\epsilon_t$. It is, furthermore, shown (cf. heavy continuous line) that two-state dimers are less stable than three-state dimers $[\epsilon_{designed}(3\text{-state})<\epsilon_{designed}(2\text{-state})]$, which in turn are again less stable than the monomer $[\epsilon(S^{36})<\epsilon_{designed}(3-\text{state})]$ which folds to the native conformation corresponding to one of the two identical halves that build the dimer [cf. Fig. 1(a)].

Since the folding properties of homodimeric sequences depend on their conformational ground state energy, it is interesting to study the energy landscape of the space of sequences for a given conformation—space that is responsible for the evolution of the corresponding dimer. For this purpose, we have calculated the entropy as a function of the bulk energy E_1 and of the interface energy E_t in the space of sequences, making use of a multicanonical algorithm [21,22], keeping fixed the conformation of Fig. 1. The results are displayed in Fig. 4. Also displayed in this figure are the absolute minima for the bulk and for the interface contacts.

As already seen from Fig. 3, it is not possible to optimize the bulk and interface at the same time. In fact, lowering ϵ_t is done at the expenses of ϵ_1 and vice versa. Furthermore, the results displayed in Fig. 4 indicate that the condition of en-

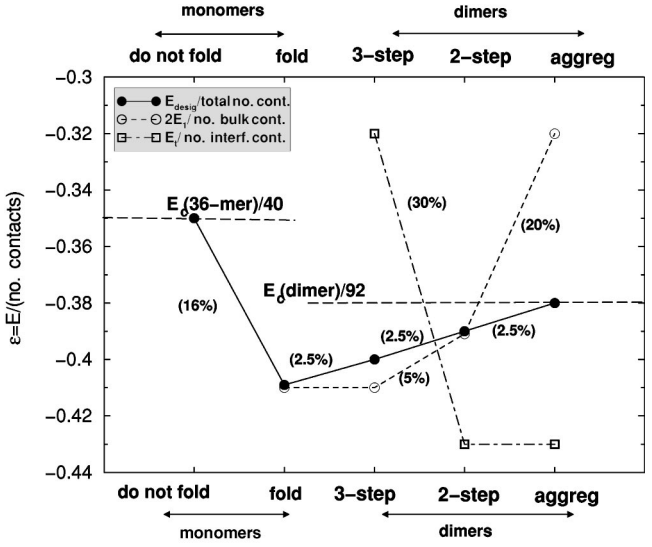

FIG. 3. Average energy per contact associated with the bulk $(2E_1/80)$, with the interface $(E_t/12)$, and with the energy of the two identical designed monomers in the native conformation $E_{desig}/92=(2E_1+E_t)/92$ of the dimer whose native conformation is shown in Fig. 1(a). Also shown are the results associated with the isolated monomer S_{36}. Also reported is the average energy per contact associated with the threshold energy E_c associated with both the dimer $(=-35)$ and the monomer $(S_{36}, E_c=-14.5)$. We display in parentheses the differential variation (in percent) of the quantities associated with two consecutive sequences from left to right.

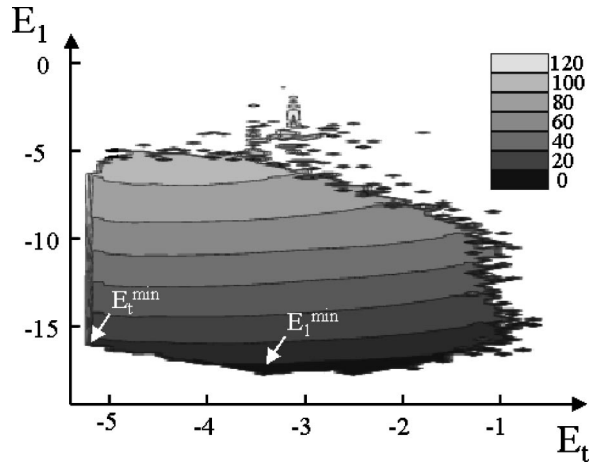

FIG. 4. The entropy of the space of sequences as a function of E_1 and E_t. The lowest energy states with respect to E_1 and E_t are indicated. The isoentropic curves that separate the gray levels correspond to variations in the entropy of 20 (in the same dimensionless units used also for the energy). The darker areas correspond to lower values of the entropy.

ergy minimum for the bulk energy ($E_1^{min}=-17.1$) implies that the interface energy is quite far from its minimum ($E_t=-3.4$). Vice versa, the minimum of the interface energy $E_t^{min}=-5.11$ is associated with a bulk energy $E_1=-16.2$. In keeping with this result, it is seen that small changes in E_1 are correlated with large variations in E_t, but not vice versa. In fact, a 2.5% change in E_1 is correlated with a 30% change in E_t, much larger than the ratio between the number of bulk and interface contacts (≈ 7). In other words, bulk contacts play a central role in the design of both three-step as well as two-step folding dimers, as testified by the fact that in both cases $\epsilon_1 < \epsilon_c$, a condition not required to be fulfilled by ϵ_t.

The exclusion relation observed between a simultaneous low value of E_1 and of E_t results in negative specific heats $C_{1t} \equiv \partial E_1/\partial \tau_t$ and $C_{t1} \equiv \partial E_t/\partial \tau_1$ at low values of τ_t and τ_1, respectively, as testified by the decrease in the energy on increasing the temperature [cf. Figs. 5(a) and 5(b)]. The regions of the (τ_1, τ_t) plane corresponding to negative specific heats C_{1t} and C_{t1} are delimited by dashed curves in Fig. 2. Two-state folding sequences lie across the $C_{1t}=0$ line, meaning that a low value of τ_t is not enough to guarantee a two-state folding character; also ϵ_1 has to be low (cf. Fig. 3). Note that $C_{1t}=0$ implies that E_1 is a minimum as a function of τ_t. In fact, sequences that are in the low-τ_t region but are in the negative-C_{1t} region aggregate (empty diamonds in Fig. 2), since the associated values of E_1 are high (cf. Fig. 3, sequence S_5). On the other hand, three-state folders are quite insensitive to the value of E_t and consequently can be found equally well on both sides of the line corresponding to $C_{t1}=0$ (i.e. also $\epsilon_t > \epsilon_c$ in Fig. 3 and solid triangles in Fig. 2).

Another interesting feature of the space of dimeric sequences is that bulk energy E_1 depends strongly on τ_1 (specific heat $C_{11} \equiv \partial E_1/\partial \tau_1$ ranging from 2 to 10) but weakly on τ_t; derivative C_{1t} being approximately a constant around 0.2 (except, of course, in the low tail of τ_t where C_{1t} is negative). This can also be seen in Fig. 3 from the fact that a

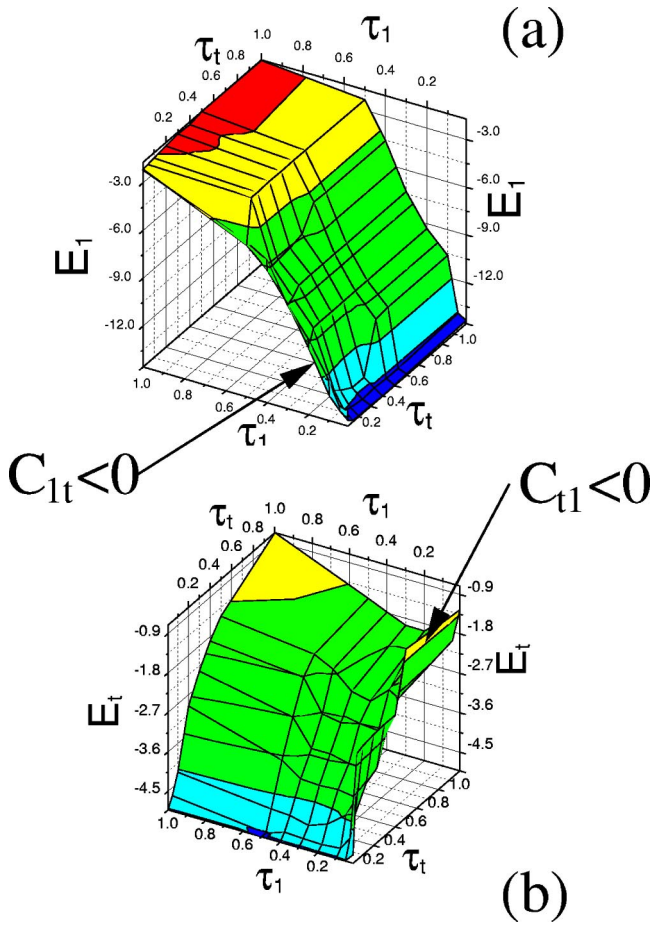

FIG. 5. The average energy E_1 (a) and E_t (b) as functions of τ_1 and τ_t.

30% change in ϵ_t (in going from sequence S_3 to sequence S_1) is related to a modest change in ϵ_1. On the other hand, the interface energy E_t, although depending strongly only on τ_t, is more sensitive to both evolutionary temperatures, as is clear from the fact that $C_{tt} \equiv \partial E_t/\partial \tau_t$ ranges from 2 to 7.5, while C_{t1} ranges from 1 to 2 (except in the low tail of τ_1).

Summing up, bulk contacts play a leading role in determining the thermodynamical properties of the designed sequence. This is not only because there are more bulk (40) than interface (12) contacts. In fact, the specific heat *per contact* is $C_{11}/40 \approx 0.2$ for the bulk and $C_{tt}/12 \approx 0.8$ for the interface. Consequently, each interface contact is more sensitive to the evolutive pressure (i.e., the evolutive temperature) exerted on it than a bulk contact, not because they are less but because of their different topology (see the following Section).

Moreover, a relation similar to the fluctuation-dissipation theorem holds:

$$\langle \epsilon_1^2 \rangle - \langle \epsilon_1 \rangle^2 = \tau_1^2 C_{11}/40,$$

$$\langle \epsilon_t^2 \rangle - \langle \epsilon_t \rangle^2 = \tau_t^2 C_{tt}/12, \tag{5}$$

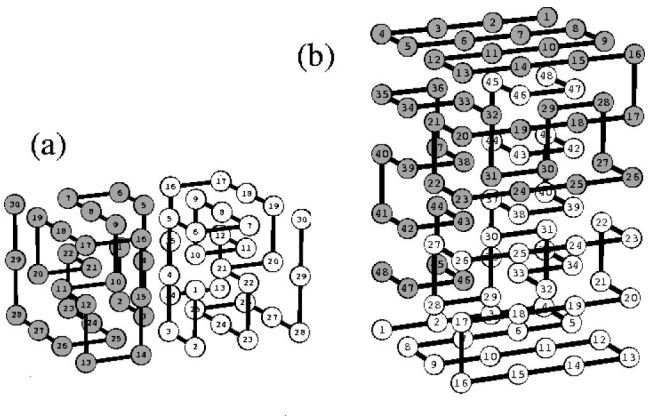

FIG. 6. Two dimeric native structures used in the present calculations, whose interfaces have some degree of intertwining.

which links the energy fluctuations (left hand side) to the specific heat (right hand side). This is a consequence of the fact that the distribution probability (1) is separable in E_1 and E_t. Consequently, for values of τ_1 and τ_t similar and in keeping that $C_{tt}/12 > C_{11}/40$, one can conclude that each interface contact fluctuates more than a bulk contact.

Let us conclude this section by noting that the monomer-monomer interface of dimers sometimes is not flat as displayed in Fig. 1 and may show some structural interpenetration in its native state (cf., for example, the case of P22 Arc repressor). Thus, to test the generality of our results, we have optimized, to various degrees, sequences on the two conformations displayed in Fig. 6 and have studied their folding properties. These conformations display a complicated, nonflat interface (compatible with the lattice model) and, moreover, a different ratio between the bulk and interface contacts than the conformation studied previously. The results are listed in Table I and, although not constituting an exhaustive sampling of sequence space as those associated with the conformation shown in Fig. 1, are in overall agreement with the findings displayed in Fig. 2. On the basis of these results, we find the fact that even dimers with a nonflat interface, display a marked difference between the bulk, built out of a complicated network of contacts, and the interface, where this feature is essentially absent (see the following section).

TABLE I. The result of dynamical simulations performed with sequences optimized at evolutionary temperatures τ_1 and τ_t on the conformations displayed in Fig. 6. In the first column, we indicate whether the results refer to conformation (a) or (b) of this figure. The fourth and fifth columns display the bulk and the interface energies of the native state, respectively. The last column indicates the kind of dynamics observed from MC simulations (see text).

Conformation	τ_1	τ_t	E_1	E_t	Kinetics
(a)	0.01	0.01	−11.04	−9.47	2-state
(a)	0.01	0.02	−11.12	−8.69	2-state
(a)	0.01	0.10	−13.37	−4.23	3-state
(a)	0.10	0.10	−12.18	−2.08	3-state
(a)	0.10	0.01	−7.86	−10.39	Aggreg
(b)	0.01	0.01	−17.02	−19.26	2-state
(b)	0.01	0.10	−22.48	−7.28	3-state
(b)	0.10	0.01	−13.13	−21.62	Aggreg

IV. CONSEQUENCES OF THE DIFFERENT TOPOLOGY OF BULK AND INTERFACE CONTACTS

The two main properties the space of sequences of dimeric proteins display are as follows: (a) It is not possible to optimize both the bulk and the interface contacts at the same time and (b) the thermodynamical quantities characterizing the designed sequences are more sensitive to the evolutionary pressure associated with τ_1 (bulk) than to that associated with τ_t (interface).

Property (a) is typical of all systems with disordered interactions. As in spin glasses, the disorder of the matrix elements makes it impossible to optimize all parts of the system at the same time (cf., e.g., [10,27]). This also happens in the case of monomeric, single-domain proteins, although one does not, as a rule, discriminate between the different contacts, imposing the same selective pressure (i.e., evolutionary temperature) on all the residues.

The asymmetry between the behavior of the bulk and of the interface is due not only to the fact that there are more bulk than interface contacts, but also because the topology of such contacts is very different. In fact, the bulk is a three-dimensional system, composed of a rich network of interconnections [cf. Fig. 1(b)], so that a residue can be correlated with other residues leading to long range order. On the other hand, the interface is composed of contacts that are essentially independent of each other, although they are not independent of the bulk contacts [cf. Fig. 1(c)]. This implies that, in principle, it is easier for an interface contact to be at a low energy than for a bulk contact, in keeping with the fact that it is easier to optimize an uncorrelated system than a correlated one. But it also means that the interface is less stable then the bulk, the associated native contact energies displaying much stronger variations than that associated with bulk contacts.

The basic difference existing between the bulk and interface contacts can be further clarified by comparing the thermodynamics of a typical lattice model protein (in this case a 36-mer) and a system composed of the same number of contacts, but placed, in the lattice, independently of each other [28]. In Fig. 7 is displayed the average energy for the protein contacts and for the independent contact system (solid and dashed curves, respectively). The average energy, sum of uncorrelated contributions, can reach an average value that is, at any design temperature (evolutionary pressure), lower than the energy of the bulk. The shape of the energy function is well described by the predictions of the random energy model [28], i.e., $\langle E \rangle \sim -T^{-1}$ (dotted curve in Fig. 7). On the other hand, the specific heat of the independent contact system (i.e. the slope of the dashed curve) is always larger than that of the bulk, indicating that the fluctuations in the independent bond system, which, through the fluctuation-dissipation theorem, are proportional to the specific heat, are larger than in the bulk. In other words, independent contacts can reach lower energies, but are more unstable than correlated contacts.

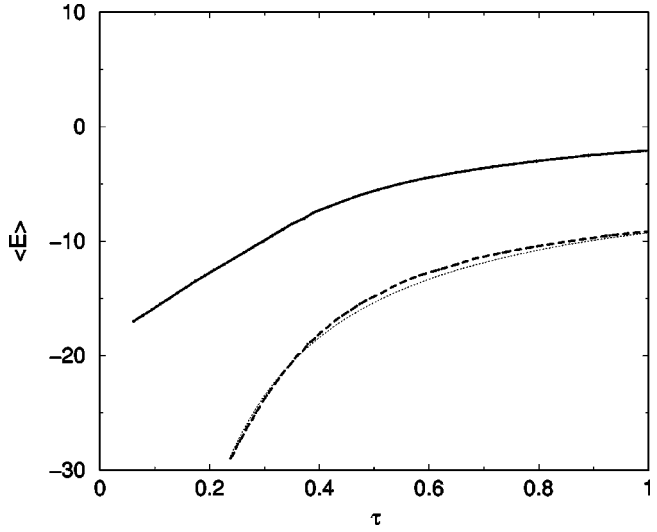

FIG. 7. The average energy as a function of evolutionary temperature for a monomeric model protein of length 36 (cf. Ref. [6], solid line) and for a system of 36 independent bonds (dashed line). The monomeric protein displays the random energy behavior $\langle E \rangle \sim \tau^{-1}$ at high evolutionary temperature and a linear behavioral low temperature, while the independent bond system displays an almost perfect random energy behavior (dotted line).

The different thermodynamics associated with the interface and with bulk contacts is also reflected by the different conservation patterns for the amino acids lying on the interface and in the bulk. Because three-state dimers concentrate their energy in the highly interconnected bulk, they display a conservation pattern typical of monomeric proteins in which few sites are highly conserved while the remaining sites can mutate more freely [12]. If one calculates for each site its sequence entropy $S(i) = -\sum_{\sigma=1}^{20} p_i(\sigma) \ln p_i(\sigma)$, where $p_i(\sigma)$ is the probability to find the amino acid of kind σ in the ith site, then the resulting distribution of $S(i)$ displays a low-S tail [cf. Fig. 8(b)].

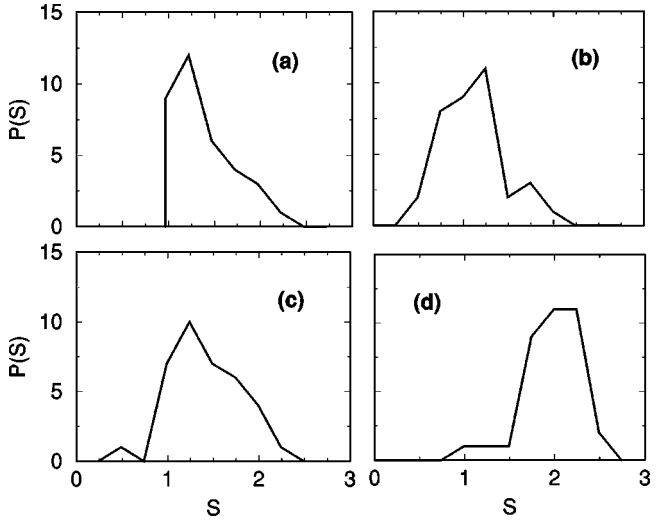

FIG. 8. The distribution $P(S)$ of entropy per site, calculated for (a) sequence 1 (two-state dimer), (b) sequence S_3 (three-state dimer), (c) sequence S_5 (aggregation), and (d) sequence S_{36} (monomer [12]).

On the contrary, in the case of two-state dimers, energy is concentrated on the interface. Since the interface contacts are independent on each other, none of them is privileged, so that the degree of conservation is more uniformly distributed among a large number of amino acids than before. The resulting distribution of $S(i)$ is shown in Fig. 8(a) and displays a sharper behavior than that associated with three-state dimers. The shapes of the two distributions agree with those found from the comparison of the sequences of real proteins (see Ref. [4]). Note that real proteins also show conservation due to functional purposes, a fact that is absent in the model proteins (which are selected only for their folding properties). Furthermore, the distributions of entropy we found are in overall agreement with the findings by Grishin and Phillips, who analyze the conservation of residues on the surface of five two-state dimeric enzymes and find no sign of any larger conservation on the interface than in the bulk [29].

The different topologies also imply that while the stabilization energy of three-state dimers is concentrated in few sites buried in the bulk and a mutation of one of these ''hot'' sites causes misfolding of the protein [12], in two-state dimers the energy is distributed more evenly on the interface, so that its sites are more tolerant to mutations. This allows two-state dimers to build active sites on the interface, for which purpose the protein has to mutate stabilizing residues with residues that perform other biological tasks.

V. FOLDING AND AGGREGATION

The folding mechanism of three-state homodimers is, within the lattice model, essentially the same as that of monomeric proteins, with the additional association step. First, LES are stabilized by strongly interacting residues, which are close along the chain (the conserved residues discussed above). When the LES assemble together to form the folding nucleus, the protein folds to its monomeric native state [6,30]. The time limiting step is the association of the two monomers into the dimers, which is controlled both by the diffusion constant and by the stability of the interface.

The behavior of two-state dimers is different. First the interface is built, which is the time limiting step, and then the rest of the protein folds around the interface. This is again a nucleation event, which causes a sudden transition from the unfolded state to the dimeric native conformation, but involves a larger number of residues than in the case of three-state dimers. This mechanism is compatible with the finding that two-state dimers, such as arc repressor [31], display small φ values (which is defined as the relative change in free energy between the native conformation and the transition state ensemble upon mutation [32]). In fact, since the transition states are determined by the association of the dimers and the residues at the interface share evenly their stabilization energy, the free energy difference between native and transition states is also distributed among a large number of residues. In other words, none of the interface residues has a leading role in the formation of the interface.

Sequences for which the strong evolutionary pressure on the interface is not balanced by a strong pressure on the bulk displays specific aggregation. The term specific means that,

although unstructured, the aggregate display some recursive interactions, typically between residues belonging to the interface. For example, in the case of sequence S_5 listed above, that between residue 3 of one monomer and 6 of the other monomer. In the design of this sequence, the low value of τ_t (strong evolutionary pressure) selects for the interface a subset of the 20 kinds of amino acids with quite low average contact energy (their average value being $\overline{B'}=-0.12$, as compared to a zero value for the MJ contact energies, with standard deviation $\sigma'_B=0.32$, to be compared with $\sigma=0.3$ for the MJ contact energies [33]). A concentration of such strongly interacting residues makes it easy the assembly of the interface in a number of ways which are different from that of the native conformation, in a similar way as if a monomeric protein was composed only of strongly interacting residues, it would display a myriad of low energy conformations competing with the native one. To be more precise, the native interface ($E_t=-5.11$) has to compete with other conformations that the first twelve residues of each chain can assume, conformations that have energies of the order of $12\overline{B'}-12\sigma'_B(2\ln\gamma)^{1/2}=-6.2$ (again evaluated in the approximation of the random energy model [28]), thus energetically more favorable than the native interface. Conversely, the bulk is not so well optimized ($E_1\sim-15$) as to reestablish the energetic balance in favor of the native state. The outcome is a globule containing the first, strongly interacting, 12 residues of each chain, surrounded by a disordered cloud made of other monomers.

This kind of aggregation is different from the aggregation of poorly designed sequences. In this case, the energy of the native state is quite high and more homogeneously distributed among the contacts, and the resulting equilibrium conformation is an ill defined clump stabilized by random interactions.

One could speculate on the fact that this ''ordered'' aggregation could be connected with the formation of fibrils. From this point of view, the destabilization of the bulk of a two-state dimer by point mutations will, as a rule, shift its location in the phase space shown in Fig. 2 from its position on the curve corresponding to $E_c=0$, into the area associated with ordered aggregation (cf. also Ref. [34]). On the other hand, to change a three-state dimer into a dimer that aggregates in an ordered way, one should mutate the interface contacts in such a way to stabilize them—a scenario that is much less likely than the previous one.

All the results described above have been found keeping fixed the size ($L=7$) of the cell that contains the system. The dependence of the behavior of the system on this size, that reflects the concentration of monomers, is described in the Appendix.

VI. CONCLUSIONS

It has been shown that, due to their different topologies, the bulk and the interface contacts contribute differently to the folding mechanism of dimers. That is, the bulk displays a rich network of contacts, while pairs of residues belonging to different monomers interact independently of each other. This difference manifests itself in the different conservation patterns expected in the case of two-state and three-state dimers—conservation patterns that are the main point of contact between model predictions and real proteins.

APPENDIX

The folding behavior dependes not only on the energetics of the sequence, but also on the monomer concentration ρ, a quantity that is reflected, within the present model, by the linear dimension L of the Wigner cell where the calculations are carried out through the expression $\rho=2/L^3$.

For $L<6$ the chains experience excluded volume violations. We found that even at $L=6$ the chains get entangled at any temperature, never finding the native conformation. This is due both to dynamical reasons, because of the lack of space available to the movements, and thermodynamical ones, since each lateral site of the native conformation interacts with those of a (virtual) neighboring dimer, interaction which was not optimized in the design process and, consequently, raises the total energy of the system above E_c.

For large L, the dimeric native state becomes unfavorable with respect to states in which the chains are separated, due to the contribution of translational entropy. In the case of sequence $S_{2\text{-}state}$, the dimeric native state ''N_2'' has to compete with the state in which the two chains are disjoint and folded in their monomeric conformation ($2N_1$). The free energy F_N of the dimeric native state is equal to its internal energy $E_N=-36.8$ (the state is unique, so its entropy is zero). The free energy of the state $2N_1$ is $2E_1-TS_{trans}$. The translational entropy S_{trans} is the logarithm of the number of different conformations in which the Wigner cell of size L can accommodate the two native conformations in such a way that they do not interact, that is, $S_{trans}=\ln[6\times4[(L+1)^3-(c+1)^3]]$, where the numeric factors take into account the different orientation of the monomers once their center of mass is fixed, the square parenthesis indicate the number of possible positions of the center of mass, in the approximation that the native monomer is a cube of length c (in the present case c is, in average, 3.3). The critical size L_c of the Wigner cell, which separates the regimes of dominance of the state N_2 rather than $2N_1$ (which corresponds to the critical concentration $\rho_c=2/L_c^3$) is obtained equating the associated free energies, which gives

$$L_c=\left[(c+1)^3+\frac{1}{24}\exp\left(-\frac{E_t}{T}\right)\right]^{1/3}-1. \tag{A1}$$

In the case of three-state dimers such as S_3, which have E_t in the range $-3/-4$, one obtains from Eq. (A1) L_c ranging from 12 to 30 at $T=0.28$, at which the folding is fastest and which regard as ''room'' temperature (cf. Ref. [6]). At values of L larger than L_c each monomer, which in the native state has an energy $E_1\ll E_c(N=36)=-14$, is stable [typically $E_c(N=36)-E_1$ is of the order of $10T$].

For two-state dimers the critical size of L_c, calculated with Eq. (A1), is of the order of 150, much larger than the critical size of three-state dimers. Moreover, even if one

chooses a low concentration (large L) such that the dimeric native state becomes unstable, each monomeric native state (N_1 state) has an energy larger than the case of sequences of kind $S_{2\text{-}state}$, i.e. of the order of few T, so that it is quite unstable, having to compete with the sea of monomeric unfolded conformations.

In the calculations of the phase diagram of Fig. 2, we have chosen $L=7$, a choice that assures stability of all sequences selected and, as discussed above, allows the movement of the chains. Even then, with this choice, the system experiences some difficulties in reaching the native conformation, due to the narrowness of the space available. For example, sequence $S_{2\text{-}state}$ can find its correct dimeric native state in 16 times out of 20. In fact, in four cases it finds a conformation with energy $E=-35.92$, corresponding to a situation in which one of the two chains (let us call them A and B) is folded (say, chain A), the monomers of chain B at the interface being in their native position, while chain B is in a (well defined) conformation that has only 40% similarity with its native structure (similarity parameter $q=0.4$). The reason for this result is to be found in the fact that chain B builds some contacts with the ''back'' of chain A, taking advantage of the periodic boundary conditions. These contacts are mostly between residues of chain B and partners that are of the right kind but belong to the wrong chain (contacts $17A-32B$, $23A-18B$, $24A-17B$, $25A-36B$, $26A-35B$, $35A-26B$), while two of them are between monomers that cannot be in contact if they belong to the same chain ($31A-32B$, $32A-33B$).

Such conformation corresponds to a metastable state that, even on slowing down the folding process, does not interfere seriously with the thermodynamics of this two-chain model. The situation is different if one considers a more realistic system built of a number of proteins comparable to the Avogadro number. In this case, the system can build a chain of dimers that gain an additional energy through this back binding. This could give rise to regular structures resembling amyloid fibrils.

[1] D. Xu, C. Tsai, and R. Nussinov, Protein Sci. **7**, 533 (1998).

[2] M.E. Milla and R.T. Sauer, Biochemistry **33**, 1125 (1994).

[3] M. Bennet, M. Phillips, and D. Eisenberg, Protein Sci. **4**, 2455 (1995).

[4] G. Tiana and R.A. Broglia, Proteins: Struct., Funct., Genet. **49**, 82 (2002).

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

[6] R.A. Broglia and G. Tiana, J. Chem. Phys. **114**, 7267 (2001).

[7] Because the bulk energy is always much larger than that associated with the interface, due to the fact that the number of native bulk contacts is much larger than the number of interface contacts, we shall refer, as a rule, to relative bulk and relative interface contact energies.

[8] N. Go, Int. J. Pept. Protein Res. **7**, 313 (1975).

[9] K.F. Lau and K. Dill, Macromolecules **22**, 3986 (1989).

[10] H. Frauenfelder and P. Wolynes, Phys. Today **47**(2), 58 (1994).

[11] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).

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

[13] H.S. Chan and K.A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).

[14] M. Karplus and E. I. Shakhnovich, in *Protein Folding*, edited by T. Creighton (Freeman, New York 1992).

[15] The most ambitious classical molecular dynamics simulations have been able to study, starting from a random conformation, a small helix bundle [16] and to follow villin for about one-tenth of its folding time [17]. Making use of distributed computing implementation (a cluster of 30 000 volunteer computers around the world) a 23-residue protein has been followed from the denatured to the folded state [18].

[16] G. Favrin, A. Irbaek, and S. Wallin, Proteins **47**, 99 (2002).

[17] Y. Duan and P.A. Kollman, Science **23**, 282 (1998).

[18] C.D. Snow *et al.*, Nature (London) **420**, 102 (2002).

[19] E.I. Shakhnovich and A.M. Gutin, Nature (London) **346**, 773 (1990).

[20] G. Tiana, R.A. Broglia, and E.I. Shakhnovich, Proteins: Struct., Funct., Genet. **39**, 244 (2000).

[21] B. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).

[22] J. Borg, Eur. Phys. J. B **29**, 481 (2002).

[23] Whether the monomers are identical or related by mirror symmetry is immaterial within the framework of the minimal model of protein folding we employ.

[24] T.E. Creighton, *Proteins* (Freeman, New York, 1993).

[25] J. Rey and J. Skolnick, Chem. Phys. **158**, 199 (1991).

[26] That is, low with respect to the only energy scale of the system, which is the standard deviation of the interaction matrix $\sigma=0.3$.

[27] M. Mezard, G. Parisi, and M.A. Virasoro, *Spin Glasses and Beyond* (World Scientific, New York, 1988).

[28] B. Derrida, Phys. Lett. **24B**, 2613 (1981).

[29] N. Grishin and M. Phillips, Protein Sci. **3**, 2455 (1994).

[30] G. Tiana and R.A. Broglia, J. Chem. Phys. **114**, 2503 (2001).

[31] E.I. Shakhnovich, Nat. Struct. Biol. **6**, 99 (1999).

[32] A.R. Fersht, *Structure and Mechanism in Protein Science* (Freeman, New York, 1999).

[33] This result also implies a quite low local value of the threshold energy $E_c'=NB'-N\sigma_B'(2\ln\gamma)^{1/2}$ (see text). In fact, assuming it to be valid for both the interface and the bulk, one would obtain (making use of $\gamma=2.2$, which for $\overline{B}=0$ and $\sigma_{=}0.3$ gives $E_c=-35$) $E_c'=-48.3$.

[34] R.A. Broglia, G. Tiana, S. Pasquali, H.E. Roman, and E. Vigezzi, Proc. Natl. Acad. Sci. U.S.A. **95**, 12 930 (1998).