Designable structures are easy to unfold

Cristiano L. Dias and Martin Grant
Physics Department, Rutherford Building, McGill University, 3600 rue University, Montréal, Québec H3A 2T8, Canada
(Received 18 October 2005; published 9 October 2006)

We study the structural stability of models of proteins for which the selected folds are unusually stable to mutation, that is, designable. A two-dimensional hydrophobic-polar lattice model was used to determine designable folds and these folds were investigated through Langevin dynamics. We find that the phase diagram of these proteins depends on their designability. In particular, highly designable folds are found to be weaker, i.e., easier to unfold, than low designable ones. We expect this to be related to protein flexibility.

DOI: 10.1103/PhysRevE.74.042902 PACS number(s): 87.15.Aa, 87.15.By, 87.15.He

While the number of different proteins exceeds 10^5, when classified in terms of structures, only of order of 10^3 families of protein folds exist [1,2]. These structural templates for amino-acid sequences can be understood [3–5] in terms of minimal microscopic models. In these models, the positions of amino acids are restricted to lattice sites, and interaction energies between residues are described by a coarse-grained model. Emergent structures are classified by their designability, the number of different amino-acid sequences that design the same structure. A few structures are highly designable, and correspond to an enormous number of sequences. These are thereby stable to amino-acid mutation, a desirable and natural feature for evolution. As well, highly designable structures are thermodynamically stable [3,6], and have proteinlike symmetry [3,4,7].

In this Brief Report we investigate the dynamical behavior of designable structures. Some calculations suggest [8,9] that sequences of amino acids which are thermodynamically stable and whose ground state are highly designable, fold faster than random sequences. Another important aspect of proteins is their reaction to forces [10]. We study proteins under shear and find a dependence of their phase diagram on designability. This diagram reveals that highly designable structures are easier to unfold than low designable structures. This result is a consequence of how strong covalent bonds in the backbone and weak bonds are distributed in designable structures. We expect this to be related to specific function, and in particular to protein flexibility.

Topologically, a large number of α helices and a lack of β sheet secondary structures [11], seems to account for the peculiar geometry of highly designable structures. Since the type of secondary structures determines how the backbone connects surface monomers and this affects the dynamics of unfolding, one would expect that highly designable structures respond differently to forces than other structures. To investigate this, we consider a hydrophobic-polar (HP) model where a protein is a chain made up of polar (P) and hydrophobic (H) amino acids. The model incorporates hydrophobicity, the main driving force for folding [12,13]. The energy of a structural sequence is given by the short-range contact interaction,

$$\mathcal{H} = \sum_{i<j} \epsilon_{i,j}[\delta(|\vec{r}_i - \vec{r}_j| - \sigma) - \delta_{j-1,i}], \tag{1}$$

where the N monomers located at spatial positions $\vec{r}_i$ are labeled by indices i and j on a two-dimensional triangular lattice, as described below. The first delta function allows only nearest-neighbors interactions at a distance σ, while the second excludes interactions between residues which are adjacent along the backbone. The interaction energy $\epsilon_{i,j}$ between monomers i and j can have three values depending on the type of monomers being binded: H-H, H-P, or P-P. These values are chosen to minimize energy when H-like (P-like) amino acids are within (on the surface of) the protein, namely, $\epsilon_{PP} > \epsilon_{HP} > \epsilon_{HH}$. To account for the segregation of different types of amino acids an additional condition is imposed: $2\epsilon_{HP} > \epsilon_{PP} + \epsilon_{HH}$. Since compact shapes have maximum contact and the lowest energy states, they are the only shapes considered for representing proteins [14]. With this simplification, the interaction energies can be shifted without changing the relative energies of a sequence when folded into different conformations. Following Li *et al.* [3], we use $\epsilon_{HH} = -2.3$, $\epsilon_{HP} = -1$, and $\epsilon_{PP} = 0$. For studying unfolding, the triangular "lattice" is created by assigning an energy for each structure through two potentials: adjacent monomers along the backbone protein interact through harmonic potentials, others by a Lennard-Jones potential,

$$V(r_{ij}) = \sum_{i=1}^{N-1} \frac{k}{2}(r_{i,i+1} - \sigma)^2 + \frac{1}{2} \sum_{\substack{j \neq i \pm 1 \\ j \neq i}} \epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \tag{2}$$

where $r_{ij} \equiv |\vec{r}_i - \vec{r}_j|$. The harmonic bonds, with spring constant k and equilibrium length σ, ensure that the backbone of the protein is preserved during the simulation. The monomers are bound by the Lennard-Jones potential, characterized by energy ϵ and the same equilibrium length σ. These Lennard-Jones bonds can be driven apart, changing the structure of the protein. A cutoff distance of 2.5σ is used. The minimal energy structure of the model in two dimensions is a triangular lattice, up to small corrections due to surface effects. Conveniently, then, the equilibrium states can be studied by simply assuming all monomers sit on the positions of a triangular lattice.

To study dynamics, we use a Langevin approach where friction and a random force act on each monomer. The intensity of the random force is given by a fluctuation-dissipation theorem. The friction force on each monomer is proportional to the relative velocity of the monomer with respect to a prescribed velocity field, which can apply a shear [15]. If the

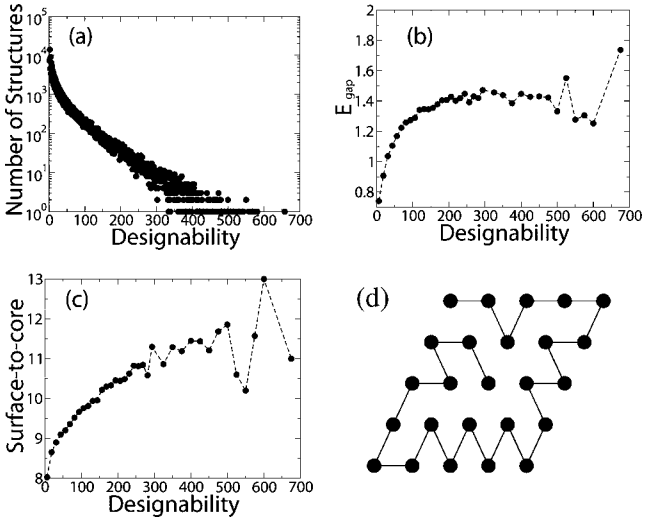

FIG. 1. (a) Histogram of designability. (b) Energy gap versus designability. (c) Number of surface to core bonds versus designability. (d) Fifth most designable structure.

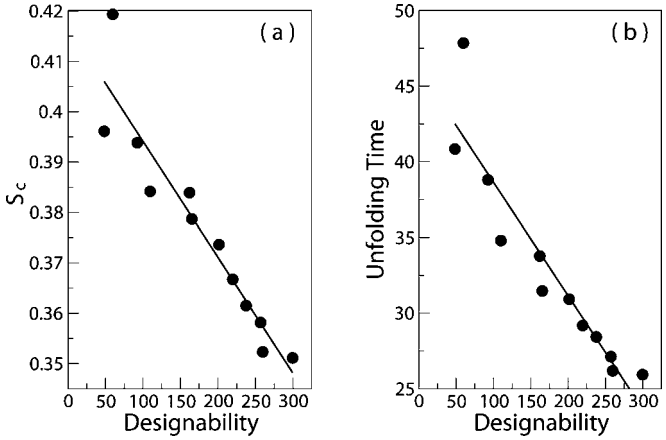

FIG. 2. (a) Dependence of S_c on designability. (b) Time required to unfold designable structures at zero shear and $T=0.5$ (in units of ϵ). Lines are just a guide to the eye.

ith monomer is located at $\vec{r}_i = x_i\hat{x} + y_i\,\hat{y}$, the prescribed velocity is $\vec{v}_{\text{fluid}}(\vec{r}_i) = Sy_i\,\hat{x}$, where S is the shear rate. The equation of motion inside the shear flow is

$$M\frac{d^2\vec{r}_i}{dt^2} = \sum_j \vec{F}(r_{ij}) - M\gamma\left(\frac{d\vec{r}_i}{dt} - \vec{v}_{\text{fluid}}(r_i)\right) + \vec{f}_i(t), \qquad (3)$$

where the sum is over all atoms inside the cutoff. Here, M is the mass of a monomer, and $\vec{F}$ is the force computed from the interacting potential. For simplicity, σ, ϵ, and M are chosen to be unity. The spring is chosen to be 5 times stiffer than the Lennard-Jones potential: $k=5(72\epsilon/\sigma^2)$. Simulations are carried out in units of the fastest atomic vibration time $\tau_0 = 2\pi\sqrt{k/M}$, and the friction constant is given a value $\gamma = (\tau_0/4)^{-1}$.

First we consider and review equilibrium structure. We study chains of 25 amino acids. Possible structures are restricted to compact self-avoiding walks on a 5×5 triangular lattice, implying 352 375 independent structures [16]. The ground state of all the 2^{25} sequences is computed and we count the number of sequences that fold uniquely into a structure. This number corresponds to the designability of a given structure. We find that 135 216 ($\sim$38%) are nondegenerate ground states of at least one sequence.

The distribution of designability for those 135 216 structures is given in Fig. 1(a). A small number of highly designable folds accommodate more than 500 sequences. These are thermodynamically stable [3]. The stability of a structure—see Fig. 1(b)—is quantified as the difference between the energy of its ground state and first excited state, E_{gap}. In this figure, E_{gap} is averaged over a given range of designabilities and plotted versus designability. Highly designable structures are seen to be more stable thermodynamically than other structures. Therefore, one can conclude that those rare structures which are highly designable, and thus stable against mutation, are also thermodynamically stable. These structures have a large number of bonds connecting surface monomers to core monomers [7,17–19]. This is shown in Fig. 1(c), where the number of bonds connecting surface to core, averaged over structures of a given range of designability, is plotted against designability. A particular example of such a highly designable structure is shown in Fig. 1(d).

Now we will quantitatively evaluate how structures with differing designabilities react to an applied shear and thermal fluctuations. Rather than simulate all 135 216 structures, we sample as follows. We study all the 1500 structures with highest designability, ranging from 200 to 700. For the more numerous structures which are less designable, we consider eight randomly chosen structures for each designability. This ensemble of 3100 structures is representative of the diversity of folds.

At zero temperature a structure only unfolds if the shear rate is greater than a critical value S_c. This critical value is a measure of structural stability to an applied force: the larger S_c is, the more stable the structure. To determine the relation between S_c and designability, we probe each structure at varying shears and different simulation times. A structure is considered to be unfolded whenever five or more bonds have broken. In Fig. 2(a), the ensemble of 3100 structures was divided into 12 bins, each containing structures with the same number (4 to 15) of surface-to-core bonds. The average designability and the average S_c of each bin is plotted in the figure. Structures which are highly designable are easier to unfold by a shear force—that is, more unstable to a shear force—than low designable structures [20].

The other extreme condition for unfolding is zero shear and high temperatures. In this case, thermal fluctuations are the mechanism responsible for unfolding. We study how the time required to unfold a structure depends on its designability at a temperature of 0.50 (in units of ϵ). In our simulations, the unfolding time τ is computed by tracking the population of folded chains. The number of chains that unfold at time t (dN/dt) is proportional to the population of folded chains $N(t)$. In this case, $N(t)=N_0\exp(-Rt)$ where R is the rate of unfolding and the characteristic unfolding time is given by the inverse of the rate $\tau=1/R$. We use 1000 copies (i.e., $N_0 = 1000$) of each structure in the simulation. The larger the

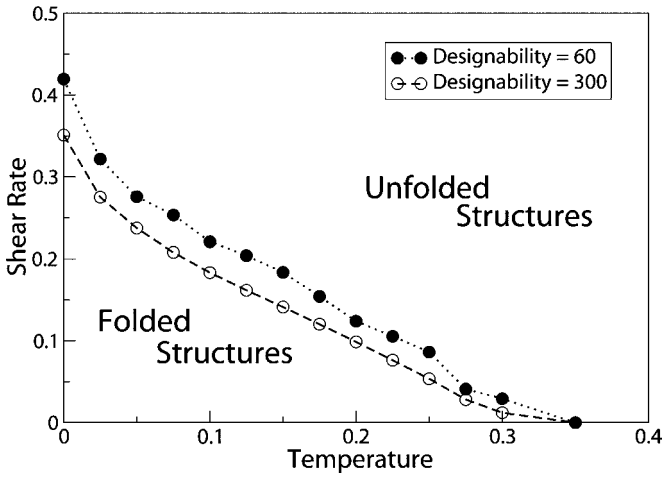

FIG. 3. Phase diagram of low and highly designable structures.

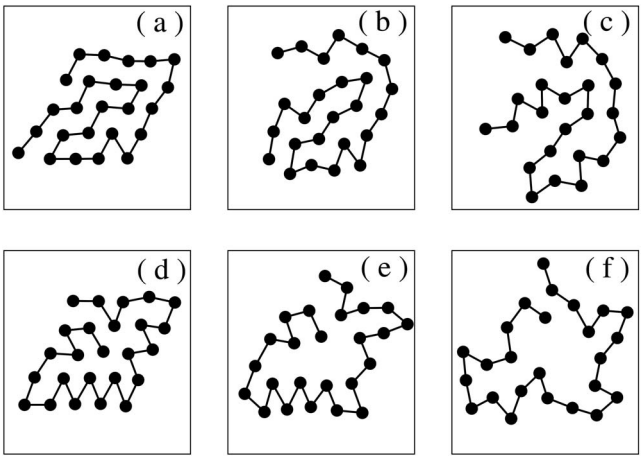

FIG. 4. Snapshot of a low [panels (a), (b), and (c)] and a highly [panels (d), (e), and (f)] designable structure during thermally induced unfolding ($S=0$ and $T=0.70$). Panels (a) and (d) show beads position at time τ_o and a time interval of $15\tau_o$ has elapsed between each panel.

unfolding time of a structure, the more stable it is to thermal fluctuations. Each point in Fig. 2(b) corresponds to the ensemble of structures having the same number of surface-to-core bonds. A clear downward trend shows that highly designable structures are less robust to thermal fluctuation: they unfold faster.

To investigate the dependence of highly designable structures on simultaneous applied shear and thermal fluctuations, the phase diagram was estimated. This diagram is constructed by computing the applied shear rate required to unfold a structure in 5000 units of time at different temperatures. This shear rate is then averaged over structures having the same number of surface-to-core bonds. Notice that the computed shear delimits two regions of the diagram: folded structures are found below this shear and unfolded structures above it. In Fig. 3 the phase diagram is shown for structures having 4 and 15 surface-to-core bonds. These two sets of structures have an average designability of 60 and 300, respectively. At any temperature, the set of structures with lower designability is more robust and require a higher shear rate to unfold. One can therefore state that high designable structures are easier to unfold than low designable ones.

It is constructive at this point to visualize the protein while it is unfolding—see Fig. 4. The upper (lower) panels of this figure correspond to the unfolding of a low (highly) designable protein fold. These simulations were performed at a temperature of 0.70 (in units of ϵ) and zero shear. Low designable folds have few surface to core bonds. As a result, many weak bonds are aligned forming substructures where monomers are correlated over long distances. For those folds, the time of unfolding is dominated by the slow unbinding of the largest substructures. In contrast, high designable folds are formed by many small substructures which are approximately of the same size. Hence it is easy to separate these substructures: only a few bonds need to rupture. This is illustrated in Fig. 4: for the low designable folds, the largest substructures is still preserved after $30\tau_o$ [panel (c)] while for the high designable folds all the small substructures have been destroyed [panel (f)].

This is in marked contrast to the relationship of designability to thermodynamic stability, namely that highly designable structures are more stable than low designable structures. The implication is that, although highly designable structures are more stable in the folded region of the phase diagram, they require less force and/or perturbation to unfold. We speculate this to be related to protein flexibility [21]: many globular proteins are stable to thermal fluctuations but undergo conformational changes (and are said to be flexible) when performing their functions. The phenomenology of this is as follows. Highly designable structures are weaker due to the large number of surface-to-core bonds they contain: as a result of this feature, protein folds contain many small substructures. These are easy to unfold since only a few bonds need to rupture in order to separate the substructures. Also, the presence in large number of surface-to-core bonds makes it difficult to transform highly designable structures into other distinct compact shapes through local rearrangements of the backbone [4]. Such a transformation would require the partial unfolding of the structure, which is unlikely in the region of the phase diagram where folded structures are at equilibrium, followed by folding into the new shape. Therefore, the presence of surface-to-core bonds might explain why high designable structures are thermodynamically stable but easier to unfold. Finally, we expect interesting insights to be obtained by expanding the model to three dimensions and modeling the solvent explicitly.

One of the authors (C.L.D.) would like to thank Anirban Sain and Markus Miettinen for many insightful discussions. This work was supported by the National Sciences and Engineering Research Council of Canada, and *le Fonds Québécois de la recherche sur la nature et les technologies*.

[1] S. E. Brenner, C. Chothia, and T. J. Hubbard, Curr. Opin. Struct. Biol. **7**, 369 (1997).
[2] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
[3] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[4] H. Li, C. Tang, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998).
[5] J. Miller, C. Zeng, N. S. Wingreen, and C. Tang, Proteins: Struct., Funct., Genet. **47**, 506 (2002).
[6] N. S. Wingreen, H. Li, and C. Tang, Polymer **45**, 699 (2004).
[7] T. Wang, J. Miller, N. S. Wingreen, C. Tang, and K. A. Dill, J. Chem. Phys. **113**, 8329 (2000).
[8] R. Melin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).
[9] R. A. Broglia, G. Tiana, H. E. Roman, E. Vigezzi, and E. Shakhnovich, Phys. Rev. Lett. **82**, 4727 (1999).
[10] C. A. Siedlecki, B. J. Lestini, K. Kottke-Marchant, S. J. Eppel, D. L. Wilson, and R. E. Marchant, Blood **88**, 2939 (1996).
[11] H. Chen, X. Zhou, and Z.-C. Ou-Yang, Phys. Rev. E **64**, 041905 (2001).
[12] W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
[13] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[14] For a study of designability on a square lattice where all self-avoiding walks with $N \leqslant 25$ are considered as possible ground states of proteins, see: Anders Irback and Carl Troein, J. Biol. Phys. **28**, 1 (2002).
[15] A. M. Maroja, F. A. Oliveira, M. Ciesla, and L. Longa, Phys. Rev. E **63**, 061801 (2001).
[16] The number of structures was computed using an algorithm based on the transition matrix method—see: Iwan Jensen, J. Stat. Mech.: Theory Exp. (2004) P10008; A. Kloczkowski and R. L. Jernigan, J. Chem. Phys. **109**, 5134 (1998).
[17] R. Helling, H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang, J. Mol. Graphics Modell. **19**, 157 (2001).
[18] C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh, and H. C. Lee, Phys. Rev. Lett. **84**, 386 (2000).
[19] C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh, J. L. Lo, and H. C. Lee, Phys. Rev. E **65**, 041923 (2002).
[20] The critical shear rate of high and low designable structures differs by up to 10% in Fig. 2. This is a large effect because thermal unfolding of structures depend exponentially upon those shear rates. When the rate is smaller than S_c, a structure can still unfold via thermal fluctuations: the time for unfolding $\tau(S) \sim \exp[E_b(S)/k_B T]$, where E_b is the energy barrier and k_B is Boltzmann's constant. Close to the critical flow, the energy landscape is dominated by the fixed point, which is a point of inflection. To lowest order in $(1-S/S_c)$ the energy barrier is $E_b(S) \propto (1-S/S_c)^{3/2}$.
[21] Dennis R. Livesay and Donald J. Jacobs, Proteins: Struct., Funct., Bioinf. **62**, 130 (2006).