Protein design depends on the size of the amino acid alphabet

Robin C. Ball*
Department of Physics, University of Warwick, Coventry CV4 7AL, England

Thomas M. A. Fink†
Morphogénèse Cellulaire et Progression Tumorale, CNRS UMR No. 144, Institute Curie, 75248 Paris Cedex 05, France and Theory of Condensed Matter, Cavendish Laboratory, Cambridge CB3 0HE, England

(Received 26 November 2001; published 19 September 2002)

We consider the design of proteins to be simultaneously thermodynamically stable in multiple independent and correlated conformations. We first show that a protein can be trained to fold to multiple independent conformations and calculate its capacity. The number of configurations that it can remember is proportional to the logarithm of the number of amino acid species A, independent of chain length. Next we investigate the recognition of correlated conformations, which we apply to funnel design around a single configuration. The maximum basin of attraction, as parametrized in our model, also depends on the number of amino acid species as $\ln A$. We argue that the extent to which the protein energy landscape can be manipulated is fixed, effecting a trade off between well breadth, well depth, and well number. This emerging picture motivates a clearer understanding of the scope and limits of protein and heteropolymer function.

DOI: 10.1103/PhysRevE.66.031902 PACS number(s): 87.14.Ee, 36.20.Ey, 87.15.Aa, 87.15.He

I. INTRODUCTION

It is believed that a stable, fast-folding protein requires an energy landscape in which the native conformation is both a deep global minimum and lies at the bottom of a basin of attraction sloping towards it [1]. These conditions are known as thermodynamic stability and kinetic accessibility, respectively. The former guarantees that, at equilibrium, a significant fraction of molecules are in the target state. The latter ensures that the protein folds in a time scale short compared to the time necessary to sample all configurations.

Not surprisingly, any practically useful method of designing artificial proteins or heteropolymers must select for stability and accessibility as well. The first satisfactory method of protein design was introduced by Shakhnovich in 1994 [2,3]: a random sequence is embedded in the target conformation and optimized over sequence space to a deep (though generally not global) minimum. At finite temperature the resulting sequence spontaneously folds to the target conformation. This strategy relies on the correlation between stability and accessibility; stable sequences are found to fold more quickly as well.

In Ref. [4], we showed that a protein can be designed to fold to multiple independent conformations, analogous to an associative memory. In this paper we generalize the notion to training to a weighted set of conformations, which proves useful in our analysis of funnel design below. In the special case of equal weights we recover the capacity result obtained in Ref. [4].

Our approach to designing multiply conforming proteins is to train to a superposition of contact maps as though it were a single map, minimizing over sequence space in the usual way [5]. We find that the number of conformations that a protein can remember is independent of the length of the protein and proportional to the logarithm of the number of amino acid species A.

While stability may be readily achieved by suppressing the energy of the sequence arranged in the target conformation, constructing a broad funnel leading towards the target has remained elusive. We next investigate the introduction of a folding funnel above the target conformation in the protein energy landscape. Our method of design rests on the technique of training to multiple targets described above. Unlike the independent configurations considered previously, here our patterns are correlated to a single target conformation.

Our approach to funnel design is to turn off all the monomer interactions (equivalent to an interacting system at infinite temperature) and to consider the dynamics by which a protein would then spontaneously *unfold* from the target state into a random ensemble. By the principle of detailed balance in equilibrium statistical mechanics, the ensemble of unfolding trajectories from the target state to random conformations is equivalent to the ensemble of folding trajectories from random configurations to the target—but of course the former ensemble is much more easily sampled. Therefore, observations of unfolding will tell us how the molecule would with least dynamical constraint fold.

We provide estimates of the unfolding contact map based on a blob model of unfolding. This is motivated by thermodynamic tractability and its basis in established polymer physics, despite its at times distorted representation of kinetics. It leads to a definite proposal as to how different stages in the unfolding contact map should be weighted in training so as to create an optimal funnel.

We find, however, that the extent of the optimal folding funnel (in terms of a relaxation length scale) is smaller than the conformational space and depends on the number of amino acid species available as $\ln A$ [6,7]. Remarkably, the

*Electronic address: r.c.ball@warwick.ac.uk
URL: http://www.phys.warwick.ac.uk/theory/
†Electronic address: fink@lps.ens.fr
URL: http://www.tcm.phy.cam.ac.uk/~tmf20/

bound on funnel size (kinetic capacity) is identical to the thermodynamic capacity derived for independent conformations. Taken together, our results suggest that the extent to which the protein energy landscape can be manipulated—whether it be by the introduction of multiple independent minima, well depth or well breadth (or a combination thereof)—is limited and proportional to the logarithm of the number of amino acid species.

II. TRAINING TO MULTIPLE CONTACT MAPS

We consider model proteins in the form of self-avoiding walks on a three-dimensional cubic lattice with nearest-neighbor interactions (lattice proteins). A protein chain consists of N amino acids (sometimes referred to as monomers), each of which is chosen from an alphabet of A amino acid species. Monomer species a and b interact according to the $A\times A$ pair potential U_{ab}. The species of the kth monomer of sequence S is denoted by S_k. For convenience of notation we also introduce the extended $N\times N$ pair potential $\tilde{U}_{ij}$, where $\tilde{U}_{ij}$ is the interaction energy between monomers i and j, that is, $\tilde{U}_{ij}=U_{S_iS_j}$.

Throughout our analysis we assume that the pair potential U has zero mean, as was found by approximating real amino acids in Ref. [8]. Otherwise, as discussed in Ref. [9], the molecule is liable to suffer the indiscriminate globular collapse characteristic of homopolymers with an attractive interaction. Our concern in this paper is the more delicate competition between the energy minima we can design and the typical deepest energy minimum found in a random copolymer.

We denote protein conformations by the contact map C, which is an $N\times N$ matrix in which $C_{ij}=1$ if S_i and S_j are nearest neighbors in the lattice and $C_{ij}=0$ otherwise [10]. We exclude contacts between monomers adjacent along the protein chain because these are fixed and cannot influence the folding dynamics. Accordingly, for compact conformations, each interior monomer is surrounded by its chain neighbors plus effective coordination number $z'=z-2$ others, where z is the lattice coordination number.

With this notation the energy of a sequence in the conformation corresponding to contact map C can be expressed as

$$E=\frac{1}{2}\sum_{ij=1}^{N} C_{ij}\tilde{U}_{ij}. \qquad (1)$$

In this section we consider training a sequence to a weighted superposition of contact maps. This is achieved by suppressing the Hamiltonian [11] of the protein in the superimposed (total) contact map in the usual way, that is, by minimizing over sequence space. We expect conformations associated with higher weights to have deeper wells. The derivation of the precise dependence follows.

The total contact map is defined by summing over the individual maps with suitable weights,

$$C_{{\rm tot}_{ij}}=\sum_{\mu=1}^{p} w_\mu C_{\mu_{ij}}, \qquad (2)$$

where w_μ is the weight associated with conformation Γ_μ. The minimum Hamiltonian associated with the total weighted contact map is

$$H_{\rm tot}^{\rm min}=\frac{1}{2}\sum_{ij=1}^{N} C_{{\rm tot}_{ij}}\tilde{U}^*_{ij}=\frac{1}{2}\sum_{ij=1}^{N}\sum_{\mu=1}^{p} w_\mu C_{\mu_{ij}}\tilde{U}^*_{ij}, \qquad (3)$$

where $\tilde{U}^*$ minimizes $H_{\rm tot}$.

We reexpress the right-hand side of Eq. (3) as a sum over the local Hamiltonian $H_{{\rm tot}_i}$, each minimized by the choice of amino acid S_i,

$$H_{\rm tot}^{\rm min}=\sum_{i=1}^{N}\min_{S_i}[H_{{\rm tot}_i}], \qquad (4)$$

where $H_{{\rm tot}_i}$ is the sum over connections to monomer i,

$$H_{{\rm tot}_i}=\frac{1}{2}\sum_{j=1}^{N}\sum_{\mu=1}^{p} w_\mu C_{\mu_{ij}}\tilde{U}_{ij}. \qquad (5)$$

The quantity $H_{{\rm tot}_i}$ is simply a weighted sum of the independent local conformational energies,

$$H_{{\rm tot}_i}=\sum_{\mu=1}^{p} w_\mu E_{\mu_i}. \qquad (6)$$

Alternatively, each $H_{{\rm tot}_i}$ may be considered as a weighted sum of $z'p/2$ independent random interaction energies sampled from the pair potential. Recalling that the distribution of bond energies U_{ab} has zero mean and calling its standard deviation σ, we approximate the distribution of $H_{{\rm tot}_i}$ by its central limit form; it is a Gaussian with zero mean and variance

$$\sigma^2_{{\rm tot}_i}=\frac{z'}{2}\sigma^2\sum_{\mu=1}^{p} w_\mu^2. \qquad (7)$$

This estimation is valid out to $|H_{{\rm tot}_i}|\sim(z'/2)\sigma\Sigma_{\mu=1}^{p}w_\mu$.

We now consider $H_{{\rm tot}_i}$ in Eq. (6) as a sum of two terms,

$$H_{{\rm tot}_i}=w_\mu E_{\mu_i}+\sum_{\nu=1,\nu\neq\mu}^{p} w_\nu E_{\nu_i}=H_{\mu_i}+H_{{\rm oth}_i}. \qquad (8)$$

Since H_{μ_i} and $H_{{\rm oth}_i}$ are independently Gaussianly distributed with variances

$$\sigma^2_{\mu_i}=\frac{z'\sigma^2}{2}w_\mu^2$$

and

$$\sigma^2_{{\rm oth}_i}=\frac{z'\sigma^2}{2}\sum_{\nu=1,\nu\neq\mu}^{p} w_\nu^2, \qquad (9)$$

the distribution of H_{μ_i} for fixed $H_{\mu_i}+H_{{\rm oth}_i}=H^{\rm min}_{{\rm tot}_i}$ reduces to

$$f(H_{\mu_i}|H^{\min}_{\mathrm{tot}_i}) \simeq c \exp\left[-\frac{\sigma^2_{\mathrm{tot}_i}}{2\sigma^2_{\mu_i}\sigma^2_{\mathrm{oth}_i}} \left(H_{\mu_i} - \frac{\sigma^2_{\mu_i}}{\sigma^2_{\mathrm{tot}_i}} H^{\min}_{\mathrm{tot}_i} \right)^2 \right], \tag{10}$$

where c is a normalizing constant and $\sigma^2_{\mathrm{tot}_i} = \sigma^2_{\mu_i} + \sigma^2_{\mathrm{oth}_i}$. The value of H_{μ_i} of maximum likelihood from Eq. (10) is given by

$$H^{\min}_{\mu_i} \simeq \frac{\sigma^2_{\mu_i}}{\sigma^2_{\mathrm{tot}_i}} H^{\min}_{\mathrm{tot}_i}, \tag{11}$$

which reduces to

$$H^{\min}_{\mu_i} = w_\mu E^{\min}_{\mu_i} \simeq \frac{w^2_\mu}{\sum_{\mu=1}^{p} w^2_\mu} H^{\min}_{\mathrm{tot}_i}. \tag{12}$$

The minimum local Hamiltonian $H^{\min}_{\mathrm{tot}_i}$ corresponds to the smallest of A samples from the distribution of H_{tot_i}. What is the minimum of M samples of a Gaussianly distributed random variable with mean 0 and standard deviation σ_X? For reasonably large M, it can be approximated [4] as

$$x^{\min} \simeq -\sqrt{2}\sigma_X \sqrt{\ln M}. \tag{13}$$

With $M = A$ and $\sigma_X = \sigma_{\mathrm{tot}_i}$, this yields

$$H^{\min}_{\mathrm{tot}_i} \simeq -\sqrt{2}\sigma_{\mathrm{tot}_i} \sqrt{\ln A}. \tag{14}$$

Substituting Eq. (14) into the right-hand side of Eq. (12) and summing over i, we find

$$E^{\min}_\mu \simeq -\sqrt{z'} N \sigma \sqrt{\ln A} \frac{w_\mu}{\left(\sum_{\mu=1}^{p} w^2_\mu \right)^{1/2}}. \tag{15}$$

This establishes how the minimized Hamiltonian distributes over the individual weighted configurations. For the special case of equal weights it reduces to

$$E^{\min}_\mu \simeq -\sqrt{\frac{z'}{p}} N \sigma \sqrt{\ln A}, \tag{16}$$

which is the result derived in Ref. [4].

III. HOW MANY CONFORMATIONS CAN A PROTEIN REMEMBER?

For a protein to fold to a target conformation Γ_μ, it is necessary that the energy of the sequence in that conformation, $E^{\min}_\mu$, be below the energy of the same sequence in conformations elsewhere. Since the target conformation is not correlated to conformations far away, distant fluctuations in the energy landscape are statistically identical to those of a random protein sequence, or copolymer. Therefore, we require that the target well be deeper than the global minimum fluctuations of a random copolymer, that is,

$$E^{\min}_\mu < E^{\min}_{\mathrm{cp}}. \tag{17}$$

In order for a protein to fold to multiple conformations, it is necessary that all of the target wells lie below $E^{\min}_{\mathrm{cp}}$.

The left-hand side of Eq. (17) is given by Eq. (16). Here we estimate the typical minimum copolymer energy $E^{\min}_{\mathrm{cp}}$. The energy of a fixed random sequence S_{cp} folded to its ground state conformation is

$$E^{\min}_{\mathrm{cp}} = \min_C \left[\frac{1}{2} \sum_{ij=1}^{N} C_{ij} \tilde{U}_{\mathrm{cp}_{ij}} \right] = \frac{1}{2} \sum_{ij=1}^{N} C^*_{ij} \tilde{U}_{\mathrm{cp}_{ij}}, \tag{18}$$

where minimization is over all C corresponding to valid conformations and C^* minimizes E_{cp}.

Since each row or column of the contact map C has z' bonds, the quantity E_{cp} from Eq. (18) (before minimization) is the sum of $z'N/2$ bonds. These contact energies are uncorrelated and may be considered as random since the extended pair potential $\tilde{U}$ of the copolymer from Eq. (18) is untrained. We therefore approximate the distribution of E_{cp} as a Gaussian $f(E_{\mathrm{cp}})$ with $\sigma^2_{\mathrm{cp}} = (z'N/2)\sigma^2$, in accordance with the central limit theorem. This estimation is valid out to $|E_{\mathrm{cp}}|$ of order $(z'N/2)\sigma$. Since the number of compact conformations of an N-mer grows as κ^N, where $\kappa \simeq 1.85$ on a cubic lattice [12], the ground state energy $E^{\min}_{\mathrm{cp}}$ is the minimum of κ^N samples of $f(E_{\mathrm{cp}})$. (A broader treatment of Hamiltonian walks, including the possibility of anomalous N dependence for less coordinated lattices, can be found in Ref. [13].) Substituting $\sigma^2_{\mathrm{cp}} = (z'N/2)\sigma^2$ for σ_X and κ^N for M in Eq. (13), we find

$$E^{\min}_{\mathrm{cp}} \simeq -\sqrt{z'} N \sigma \sqrt{\ln \kappa}. \tag{19}$$

Inserting Eqs. (16) and (19) into the condition for folding Eq. (17), the bound on the number of targets p that a protein can remember is found to be

$$p_{\max} \simeq \frac{\ln A}{\ln \kappa}. \tag{20}$$

The capacity may also be derived by information theoretic considerations. Imagine the transmission of a conformation encoded as a sequence. The conformation is decoded by constructing the sequence and allowing it to fold, either *in vivo* or via computer simulation.

How much information is transmitted? Since a protein sequence is simply an N digit number in base A, the information sent (in bits) is $\log_2$ of the number of possible sequences, that is,

$$I_{\mathrm{trans}} = N \log_2 A. \tag{21}$$

The $A \times A$ pair potential U is part of the decoding apparatus (encoded by the laws of physics or included in the folding algorithm) and need not be transmitted each time a conformation is sent.

The information received can be similarly calculated. If we label the conformations $1,2,\ldots,\kappa^N$, then decoding a single conformation amounts to receiving $\log_2\kappa^N$ bits. Since the target conformations are (by assumption) independent, the information contained in the p lowest-energy conformations grows linearly with p, that is,

$$I_{\mathrm{rec}} = pN\log_2\kappa. \tag{22}$$

The information received must be less than or equal to the information transmitted, that is, $I_{\mathrm{rec}} \leqslant I_{\mathrm{trans}}$. Accordingly, $pN\log_2\kappa \leqslant N\log_2 A$, from which it follows that

$$p_{\max} \simeq \frac{\ln A}{\ln \kappa}, \tag{23}$$

in agreement with Eq. (20). We discuss the implications of our capacity result in Sec. VI. In what follows we focus on the design of a broad kinetic folding funnel.

IV. BLOB MODEL OF UNFOLDING

It is a well known trend in polymer physics that the larger scale features of molecular conformations have systematically longer relaxation times. For example, for noninteracting chains with simple kink-jump dynamics, a subsection of g monomer units has relaxation time $\tau(g)$ proportional to g^2 [14]. On this basis we assume that after time t, a spontaneously unfolding polymer will have equilibrated locally up to scale g, such that $\tau(g)=t$, but still reflect the folded conformation on larger scales.

This blob view of proteins, that time scales relate uniformly to length scales, is of course a particular and simplified outlook, motivated by its tractability. Complications that we do not address here include spatially localized nucleation events and specific configurational bottlenecks. Nevertheless, it allows us to make some quantitative predictions about the limits of the basin of attraction, which has long proved to be evasive.

The folded protein, which we assume to be compact and associate with $g=1$, consists of N single monomer blobs. The contact map $C(1)$ has z' nonzero entries in each row and column, $z'N$ nonzero entries in total.

For the state unfolded up to length scale g, the protein may be thought of as a chain of N/g blobs, folded to its coarse-grained original conformation. Accordingly, the contact map $C(g)$ has N/g intrablob blocks along the diagonal and $z'N/g$ interblob blocks corresponding to nearest-neighbor blobs (not along the backbone). Scaling theories for polymer configurations with excluded volume would imply that the average total number of contacts between two neighboring blobs be of order unity. Averaging over an ensemble of conformations at constant g, this requires that each of the g^2 entries for each blob be of order $1/g^2$.

The total number of conformations (compact or otherwise) available to a protein grows as $\sim\tilde{\kappa}^N$ [14] (not to be confused with $\kappa \simeq 1.85$ for compact structures only); this becomes $\tilde{\kappa}^{N/g}$ for a chain of N/g blobs. Since the product of the internal and external conformational freedoms of a partially relaxed protein must equal $\tilde{\kappa}^N$, a protein relaxed to length scale g can be estimated to take on $\tilde{\kappa}^{[N-(N/g)]}$ configurations. It follows that the entropy gained in folding from a denatured configuration down to a conformation relaxed to length scale g is

$$S(g) = -k_B\frac{N}{g}\ln\tilde{\kappa}. \tag{24}$$

V. HOW BIG IS THE OPTIMAL FOLDING FUNNEL?

While an energy minimum significantly below the minimum copolymer energy ensures thermodynamic stability of the target conformation, rapid convergence requires a funnel of kinetic pathways sloping towards the target. The widest possible funnel is that which least constrains the dynamics, which we propose is given by the conformations sampled in unfolding via the blob model. We thus consider combining the contact maps from different times (and values of g) of a noninteracting, spontaneously unfolding compact conformation with weights $w(g)$,

$$C_{\mathrm{tot}_{ij}} = \sum_{\ln g=1}^{\ln N} w(g)C_{ij}(g). \tag{25}$$

The minimum Hamiltonian associated with the total contact map then appears as

$$H_{\mathrm{tot}}^{\min} = \frac{1}{2}\sum_{ij=1}^{N}\sum_{\ln g=1}^{\ln N} w(g)C_{ij}(g)\tilde{U}_{ij}^{*}, \tag{26}$$

analogous to Eq. (3). The total Hamiltonian associated with monomer i is the sum of the individual local Hamiltonians evaluated at different values of g,

$$H_{\mathrm{tot}_i}^{\min} = \sum_{\ln g=1}^{\ln N} H_i^{\min}(g), \tag{27}$$

where $H(g)=w(g)E(g)$. In accordance with our previous calculation, we require $\sigma^2_{\mathrm{tot}_i}$. We first estimate the variance in the choice of $H(g)$ available to a single monomer as

$$\sigma_{g_i}^2 \simeq \frac{z'g}{2}\left(\frac{w(g)}{g^2}\right)^2\sigma^2, \tag{28}$$

where $z'g/2$ is the number of contacts available to a given monomer equilibrated to scale g and $w(g)/g^2$ is the overall weighting for each one. The variance of the local energy per monomer integrated over all g is then

$$\sigma_{\mathrm{tot}_i}^2 \simeq \sum_{\ln g=1}^{\ln N}\sigma_{g_i}^2 \simeq \frac{z'\sigma^2}{2}\int_e^N \frac{dg}{g}\,g\,\frac{w^2(g)}{g^4}. \tag{29}$$

Again we wish to establish how the minimized Hamiltonian distributes over weighted configurations unfolded to length scale g. Applying the general result (11) yields

$$H_i^{\min}(g)=w(g)E_i^{\min}(g)\simeq\frac{\sigma_{g_i}^2}{\sigma_{\mathrm{tot}_i}^2}H_{\mathrm{tot}_i}^{\min}. \tag{30}$$

Substituting Eqs. (14) and (28) into Eq. (30) and summing over i, the minimum energy associated with matching the conformation at scale g can then be estimated as

$$E^{\min}(g)\simeq-\frac{z'}{\sqrt{2}}N\sigma^2\sqrt{\ln A}\frac{w(g)}{\sigma_{\mathrm{tot}_i}g^3}. \tag{31}$$

In order that the training reverse the unfolding dynamics, the required funnel must have sufficient slope, that is, $F(g)=E(g)-TS(g)<0$. Equating the two expressions $T\times$ (24) and (31) gives

$$w(g)\simeq\frac{\sqrt{2}k_BT\ln\tilde{\kappa}\;\;\sigma_{\mathrm{tot}_i}}{z'\sigma^2\sqrt{\ln A}}g^2, \tag{32}$$

and thus $w(g)\propto g^2$. Unfortunately, this form for w is inconsistent with a convergent (N independent) evaluation of σ_{tot_i} in Eq. (29). Our assumption that the training energy could reverse the unfolding dynamics does not hold for all values of g.

We consequently introduce the cutoff scale $g_{\max}$, up to which our funnel extends. Substituting Eq. (32) into Eq. (29) and reducing the domain of integration yields

$$\sigma_{\mathrm{tot}_i}^2\simeq\frac{(k_BT)^2\ln^2\tilde{\kappa}}{z'\sigma^2\ln A}\sigma_{\mathrm{tot}_i}^2\int_e^{g_{\max}}dg, \tag{33}$$

from which it follows that

$$g_{\max}\simeq\frac{z'\sigma^2\ln A}{(k_BT)^2\ln^2\tilde{\kappa}}. \tag{34}$$

The width of our funnel, as parametrized by $g_{\max}$ above, increases strongly as folding temperature T decreases. At too low a temperature, however, the coil will collapse as a random copolymer into what we presume to be a glassy state. The gain in entropy resulting from collapse will be equivalent to Eq. (24) evaluated at $g=1$ (the collapsed copolymer will be fully folded). The modest decrease in energy afforded by the minimum copolymer energy can overcome this entropic loss only at low temperature T_{cp}. Equating the minimum copolymer energy $E_{\mathrm{cp}}^{\min}$ from Eq. (19) and T_{cp} times the change in entropy Eq. $(24)|_{g=1}$ leads to

$$k_BT_{\mathrm{cp}}\simeq\sigma\frac{\sqrt{z'\ln\kappa}}{\ln\tilde{\kappa}}, \tag{35}$$

and hence at $T\simeq T_{\mathrm{cp}}$,

$$g_{\max}\simeq\frac{\ln A}{\ln\kappa}, \tag{36}$$

which is identical to the form of $p_{\max}$ derived in Sec. III.

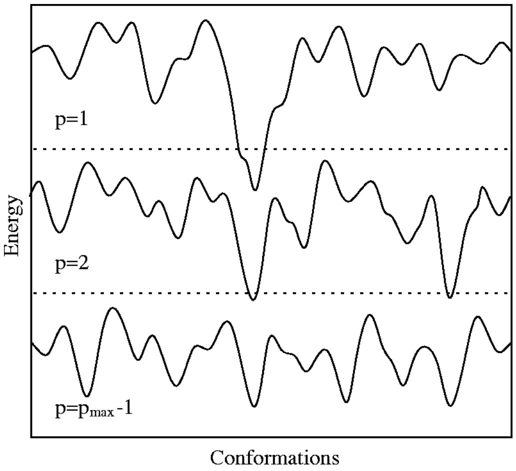

FIG. 1. Energy landscapes of sequences trained to be thermodynamically stable in a one, two, and $p_{\max}-1$ target conformations. As the number of targets increases, the depth to which the target wells can be trained diminishes. At $p=p_{\max}$, the wells are lost among nearby fluctuations.

VI. DISCUSSION OF CAPACITIES

The number of conformations that a protein can remember (thermodynamic capacity) was derived first by energetic arguments and second via information theory; in both cases we found $p_{\max}\simeq\ln A/\ln\kappa$. A homopolymer $(A=1)$ can be trained to be thermodynamically stable in 0 conformations, as expected. Binary H-P models can typically be trained to recognize 1 conformation, whereas for a protein constructed from a 20 amino acid set $p_{\max}\simeq 5$. What happens to the protein energy landscape as the number of targets is increased? As $p\rightarrow p_{\max}$, the typical depth of the target wells diminishes such that, at $p=p_{\max}$, the wells become lost in fluctuations elsewhere (Fig. 1); the targets cease to be global minima.

Insight into our thermodynamic capacity result may be gleaned from associated neural networks (ANNs), whose capacities increase linearly with the number of neurons n [15]. In both ANNs and proteins the information contained in each memory—patterns and conformations—is proportional to n and N, respectively. Unlike proteins, which can be imagined as locally connected networks, ANNs are globally connected—each of the n neurons is bonded to $n-1$ others. Since capacity is proportional to the total number of connections divided by the information contained in each pattern, the capacity of an associative memory grows linearly with n whereas proteins possess constant capacity. Perhaps more surprising is that the answer is governed by the number of amino acid species A rather than the effective coordination number z'.

That the bound on the folding funnel $g_{\max}$ is less than N implies that the extent of the achievable folding funnel (kinetic capacity) is less than the conformational space of the protein. Folding at finite temperature cannot be made as direct as unfolding at infinite temperature. The cutoff $g_{\max}$ is the length scale of the structure below which the energy landscape corresponding to the trained sequence is characterized by a funnel. Above $g_{\max}$, the protein must organize itself into the desired (coarse-grained) conformation without

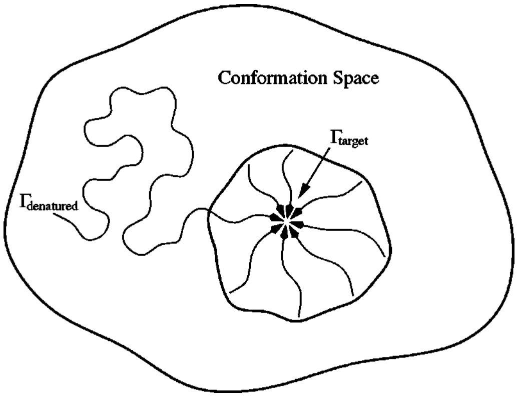

FIG. 2. Folding in the presence of a funnel. The denatured protein wanders through conformation space until it matches the target structure coarse grained to length scale $g_{\max}$, after which the funnel quickly guides the protein towards the target.

the help of kinetic guidance, that is, it must traverse an effective copolymer landscape (Fig. 2). What happens to the protein energy landscape upon increasing the width of the funnel? As $g \rightarrow g_{\max}$, the slope of the funnel becomes sufficiently shallow such that, at $g = g_{\max}$, the decrease in energy no longer overcomes the loss of entropy (Fig. 3); the well ceases to be a free energy minimum.

Consider the protein as a sequence of $N/g_{\max}$ blobs, each of size $g_{\max}$. The benefit of the funnel is realized once the chain of blobs folds to its coarse-grained target state. Assuming this statistical bottleneck to be the rate determining step, the time necessary for the protein to fold is reduced by the factor $\kappa^{-(1-1/g_{\max})N}$, which is significant even for small values of $g_{\max}$.

VII. ABILITY TO MANIPULATE THE ENERGY LANDSCAPE IS LIMITED

In both the thermodynamic (deep wells) and kinetic (broad funnel) contexts, the extent to which the protein energy landscape can be manipulated is limited by $\ln A/\ln \kappa$, where A is the number of amino acid species and κ is the compact conformational freedom per monomer. Like squeezing one end of a balloon at the expense of inflating the other, further deformation of the energy landscape is counterbalanced by its relaxation elsewhere.

The agreement between the bounds on protein memory, on the one hand, and the basin of attraction, on the other, was unexpected. Taken together, these results suggest that the engineering of proteins and heteropolymers is constrained by a fixed budget. The finite freedom in the sequence can be invested in various attributes: in well number, well breadth, and well depth. A reduction in expenditure in one allows increased investment in another.

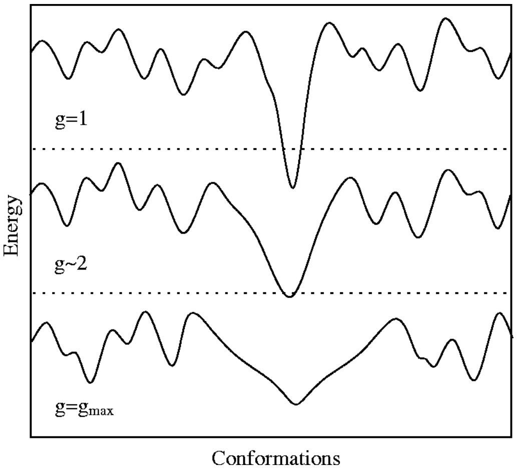

FIG. 3. Energy landscapes of sequences trained to have increasingly broad funnels. Maximizing stability (top) corresponds to a deep, narrow well. As the length scale g to which the funnel extends increases, the depth of the target well is reduced; at $g = g_{\max}$, the slope of the funnel is no longer sufficient to provide a free energy minimum (bottom).

In particular, our results suggest that thermodynamic stability and kinetic accessibility, while correlated over a significant region, are in conflict near the extremes of either; maximally stable sequences are not the fastest folding and the fastest folders are not the most stable. (We presented preliminary evidence to this end in Ref. [16]). Accordingly, a thermodynamically oriented sequence design does not select for the fastest-folding proteins and a reduction in stability admits the possibility of increased accessibility. If nature has designed proteins to fold as quickly as possible, we would expect only marginal stability in the native conformation. The preceding premise might be established by observation of normal and mutated naturally occurring proteins.

Notably, the bound on manipulating the energy landscape is independent of protein length; the diversity of protein function grows with alphabet size only. The large (relative to κ) amino acid alphabet found in nature is crucial to the variety of protein function within the cell or in multicellular organisms. To the extent that heteropolymer models are intended to provide insight into proteins, their alphabet sizes should reflect this. Elementary representations, such as the frequently studied *H-P* models, are not able to effect the thermodynamic and kinetic diversity possible with larger alphabets.

Perhaps most interesting is the increased scope for protein and heteropolymer function. The discovery that prions fold to multiple conformations [17] has extended our notion of heteropolymer behavior beyond familiar protein collapse. We have presented arguments that the energy landscape may, within limits, be tailored to effect other important functions. Further discovery of protein mechanisms should prove fascinating.

[1] Ken A. Dill and Sun Chan, Nat. Struct. Biol. **4**, 10 (1997).
[2] E.I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[3] A.M. Gutin, V.I. Abkevich, and E.I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **92**, 1282 (1995).

[4] Thomas M.A. Fink and Robin C. Ball, Phys. Rev. Lett. **87**, 198103 (2001).

[5] V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, Proteins: Struct., Funct., Genet. **31**, 335 (1998).

[6] Thomas M. A. Fink, Ph.D. thesis, University of Cambridge, 1998.

[7] The influence of alphabet size on the folding performance of *untrained* sequences is considered in J.-R. Garel, T. Garel, and H. Orland, J. Phys. (France) **50**, 3067 (1989).

[8] S. Miyazawa and R. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[9] T. Teramoto and F. Yonezawa, Int. J. Mod. Phys. B **14**, 621 (2000), and references therein.

[10] It has been argued that the native conformation of some real proteins cannot be predicted by using pairwise contact potentials [Michele Vendruscolo and Eytan Domany, J. Chem. Phys. **109**, 11 101 (1998)]. We do not address this here.

[11] When C in Eq. (1) is a single contact map, we refer to the quantity on the right as an energy; when C is a sum over multiple contact maps, the right-hand side is called a Hamiltonian.

[12] Vijay S. Pande *et al.*, J. Phys. A **27**, 6231 (1994).

[13] Manfred Gordon, P. Kapadia, and A. Malakis, J. Phys. A **9**, 751 (1976).

[14] Pierre-Gilles de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).

[15] D. J. Amit, *Modeling Brain Function* (Cambridge University Press, Cambridge, UK, 1989).

[16] Thomas M. Fink and Robin C. Ball, Physica D **107**, 199 (1997).

[17] Stanley B. Prusiner, Proc. Natl. Acad. Sci. U.S.A. **95**, 13 363 (1998).