How Many Conformations Can a Protein Remember?

Thomas M. A. Fink* and Robin C. Ball†

Theory of Condensed Matter, Cavendish Laboratory, Cambridge CB3 0HE, United Kingdom

(Received 1 May 2000; published 23 October 2001)

We show that a protein can be trained to recognize multiple conformations, analogous to an associative memory, and provide capacity calculations based on energy fluctuations and information theory. Unlike the linear capacity of a Hopfield network, the number of conformations which can be remembered by a protein sequence depends on the size of the amino acid alphabet as $\ln A$, independent of protein length. This admits the possibility of certain proteins, such as prions, evolving to fold to independent stable conformations, as well as novel possibilities for protein and heteropolymer design.

DOI: 10.1103/PhysRevLett.87.198103 PACS numbers: 87.14.Ee, 36.20.Ey, 87.15.Aa

It is widely thought to be a design feature of real proteins that their native, biologically active state is both a deep global energy minimum and has a funnel of low energy configurations leading toward it [1]. The deep well ensures that a significant fraction of protein molecules occupy the native state at any given moment. The funnel guides the molecule to fold to its stable native conformation in a time much less than that required for it to explore all configurations, thus avoiding the so-called Levinthal paradox.

Inverse protein folding, or protein design, consists of designing a sequence of amino acids that stably and quickly folds to a desired target conformation. This process may be expressed in the context of an energy landscape, to which each sequence corresponds. For each compact conformation Γ_c, there are typically a myriad of sequences which fold to it [2]. The set of sequences which fold to Γ_c corresponds to those energy landscapes whose global minima lie above the target. Most of these will possess nominally global (shallow) minima and fold in very long rather than biological time scales [1]. Of those which are deep, and hence thermodynamically stable, fewer yet will resemble broadly sloping funnels. It is this last group of energy landscapes, and hence sequences, to which natural proteins are believed to correspond. Not surprisingly, we wish to select for similar features when engineering artificial proteins.

In this sense, protein design corresponds to choosing from the spectrum of all possible sequences a sequence whose landscape possesses the attributes we desire. Because the spectrum is finite, however, we are not free to insist on an arbitrary topography; some landscapes have wells too deep or too numerous to be practicable.

In this Letter we investigate the fundamental limit on the introduction of deep (thermodynamically stable) minima into the protein energy landscape [3]. We estimate the typical maximum depth of the ground state well in a sequence trained to fold to a unique conformation. By analogy with the theory of associative neural networks (ANNs) [4], we show how protein design can be generalized to provide recognition of several conformations rather than a single target state. We find that the number of conformations that a protein can recall is limited and calculate its capacity. Remarkably, the capacity depends not on protein length but on the number of amino acid species.

The ability of a protein sequence to encode multiple conformations has immediate implications on our understanding of prions and other multistable proteins. In his Nobel lecture [5], Prusiner concludes "The discovery that proteins may have multiple biologically active conformations may prove no less important than the implications of prions for diseases. How many different tertiary structures can [a protein] adopt? This query not only addresses the issue of the limits of prion diversity but also applies to proteins as they normally function within the cell. . . ." In addition to predicting multistable proteins, our results suggest that artificial heteropolymers may be engineered to fold to multiple targets as well. We discuss possibilities for implementing target control to this end.

Proteins as associative memories.—A lattice protein consists of a sequence S of N amino acids, or monomers, each of which can take on one of A possible species. We denote the species of the ith monomer of S by S_i, and monomers i and j interact according to the $N \times N$ extended pair potential $\tilde{U}$, where $\tilde{U}_{ij} = U_{S_i S_j}$ and U is the $A \times A$ pair potential.

Protein conformations may be represented by the contact matrix C, where $C_{ij} = 1$ if monomers i and j are nearest neighbors and 0 otherwise. Contacts between monomers adjacent along the protein chain are preserved and cannot influence the folding dynamics, so we exclude these from the contact map. For compact conformations, each interior monomer is surrounded by its chain neighbors plus z' others, where z' (the effective coordination number) is two less than z (the lattice coordination number). Contact patterns are thought to be a unique representation of compact conformations and we approximate them as independent.

Protein folding may be considered pattern recognition in as much as the protein rapidly organizes itself into the target pattern C upon entering the target basin of attraction (funnel). By analogy with pattern association, this idea may be generalized to the recognition of multiple patterns. This raises the question of how to train the sequence to recognize more than one conformation. For lattice models,

Shakhnovich and co-workers [6,7] have explored the folding of sequences designed to minimize a conformation's absolute and relative energies. The essence of the training technique is to embed the protein into the target conformation and optimize stability over sequence space; the resulting (near-optimal) sequence spontaneously folds to the target. The dilute representation of conformations by contact patterns suggests that we can superimpose p patterns without saturation [8], providing us with a total pattern to which we train in the usual way. This is essentially equivalent to the method used to select bistable 36-mers in [9].

Energy function.—The energy of a sequence in the conformation corresponding to contact map C may be conveniently expressed:

$$E = \frac{1}{2}\sum_{ij=1}^{N} C_{ij}\tilde{U}_{ij}\,. \tag{1}$$

For a sequence trained to have minimal energy in conformation Γ_μ, the energy appears as

$$E_\mu^{\min} = \min_{\tilde{U}}\left[\frac{1}{2}\sum_{ij=1}^{N} C_{\mu_{ij}}\tilde{U}_{ij}\right] = \frac{1}{2}\sum_{ij=1}^{N} C_{\mu_{ij}}\tilde{U}_{ij}^{*}\,, \tag{2}$$

where minimization is over all $\tilde{U}$ corresponding to valid sequences and $\tilde{U}^*$ minimizes E_μ. The energy of a fixed sequence S_ν folded to its ground state conformation is

$$E_\nu^{\min} \equiv E_{\rm cp}^{\min} = \min_{C}\left[\frac{1}{2}\sum_{ij=1}^{N} C_{ij}\tilde{U}_{\nu_{ij}}\right] = \frac{1}{2}\sum_{ij=1}^{N} C_{ij}^{*}\tilde{U}_{\nu_{ij}}\,, \tag{3}$$

where minimization is over all C corresponding to valid conformations and C^* minimizes E_ν. As is common usage, we refer to the quantity E_ν for an untrained sequence as the copolymer energy $E_{\rm cp}$.

Throughout this Letter, the energy of a sequence realized in a particular conformation is indicated by E, while the Hamiltonian with which a sequence is trained (generally the linear combination of the energies realized in a number of conformations) is denoted by H.

Capacity from energetics.—We consider the capacity of a protein, that is, the number of conformations p that we can train the sequence to make simultaneously thermodynamically stable. For a protein to fold to a single target conformation, it is necessary that the energy of the trained sequence realized in that conformation, $E_\mu^{\min}$, be below the minimum fluctuations of the energy elsewhere. Since the trained sequence is not correlated with distant conformations, energy fluctuations away from the target structure are statistically equivalent to those of a random copolymer sequence. We therefore require that the trained energy be less than the minimum energy of a random sequence, that is, $E_\mu^{\min} < E_{\rm cp}^{\min}$. Folding to a set of p conformations requires that the minimum energy of all of these lie below $E_{\rm cp}^{\min}$.

We first estimate the typical minimum copolymer energy $E_{\rm cp}^{\min}$. Recalling that each row (or column) of the contact map C has z' bonds, the quantity $E_{\rm cp}$ from (3) (before minimization) is the sum of $\frac{z'N}{2}$ bonds. Since the extended pair potential $\tilde{U}$ of the copolymer from (3) is untrained, these contact energies are uncorrelated and may be considered random. Assuming a distribution of bonds with zero mean (as is the case of that found in [10]) and standard deviation σ, we find, in accordance with the central limit theorem, that $E_{\rm cp}$ is distributed as

$$f(E_{\rm cp}) \simeq \frac{1}{\sqrt{2\pi}\,\sigma_{\rm cp}}\exp\left(-\frac{E_{\rm cp}^2}{2\sigma_{\rm cp}^2}\right), \tag{4}$$

where $\sigma_{\rm cp}^2 = \frac{z'N}{2}\sigma^2$. This estimation is valid out to $|E_{\rm cp}|$ of order $\frac{z'N}{2}\sigma$. The ground state energy $E_{\rm cp}^{\min}$ is the least of all possible samples of (4), each of which corresponds to a unique conformation. Since the number of compact conformations of an N-mer grows as κ^N, where $\kappa \simeq 1.85$ on a cubic lattice [11], the energy of the ground state is the minimum of κ^N samples of $f(E_{\rm cp})$.

What is the minimum of M samples of a random variable X distributed according to a Gaussian $g(x)$? For convenience we assume zero mean and standard deviation σ_X. The probability distribution of x being the minimum of M samples of X is given by

$$g^{\min}(x) = Mg(x)[1 - G(x)]^{M-1}, \tag{5}$$

where $G(x) = \int_{-\infty}^{x} g(x')\,dx'$ is the usual cumulative distribution. Maximizing $g^{\min}$ with respect to x yields the transcendental equation $x^{\min}[1 - G(x^{\min})] = -\sigma^2(M - 1)g(x^{\min})$, where $x^{\min}$ is the minimum of the M realizations of X. For reasonably large M, $G(x)$ is small and we estimate $x^{\min}$ as

$$x^{\min} \simeq -\sqrt{2}\,\sigma_X\sqrt{\ln M}\,. \tag{6}$$

By way of (6), we can express the ground state energy $E_{\rm cp}^{\min}$ as

$$E_{\rm cp}^{\min} \simeq -\sqrt{2}\,\sigma_{\rm cp}\sqrt{\ln(\kappa^N)} = -\sqrt{z'}\,N\sigma\sqrt{\ln\kappa}\,. \tag{7}$$

We now approximate the typical energy of a sequence optimally trained to a set of p target conformations and arranged in one of these configurations. The total contact map, to which we train by energy minimization with respect to the sequence [6], is defined as a linear superposition of the p corresponding contact maps, that is

$$C_{{\rm tot}_{ij}} = \sum_{\mu=1}^{p} C_{\mu_{ij}}\,. \tag{8}$$

The minimum Hamiltonian associated with the total contact map may then be written

$$H_{\rm tot}^{\min} = \frac{1}{2}\sum_{ij=1}^{N} C_{{\rm tot}_{ij}}\tilde{U}_{ij}^{*} = \frac{1}{2}\sum_{ij=1}^{N}\sum_{\mu=1}^{p} C_{\mu_{ij}}\tilde{U}_{ij}^{*}\,, \tag{9}$$

where here $\tilde{U}^*$ minimizes $H_{\rm tot}$. It is simply the sum of the p individual conformational energies of the sequence implied by $\tilde{U}^*$. We reexpress the right side of (9) as the

sum over i of the total energy associated with monomer i, H_{tot_i}, each minimized with respect to the choice of amino acid at monomer i,

$$H_{\mathrm{tot}}^{\mathrm{min}} = \sum_{i=1}^{N} \min_{S_i}[H_{\mathrm{tot}_i}]; \tag{10}$$

H_{tot_i} is obtained by summing over the connections to monomer i,

$$H_{\mathrm{tot}_i} = \frac{1}{2} \sum_{j=1}^{N} \sum_{\mu=1}^{p} C_{\mu_{ij}} \tilde{U}_{ij} \,. \tag{11}$$

Since C has z' bonds connecting to monomer i, each H_{tot_i} is the sum of $\frac{z'p}{2}$ random interaction energies freely chosen from the pair potential [12]. As before, we approximate the distribution of H_{tot_i} by its central limit theorem form; it is a Gaussian with variance $\sigma_{\mathrm{tot}_i}^2 = \frac{z'p}{2}\sigma^2$. This estimation is valid out to $|H_{\mathrm{tot}_i}|$ of order $\frac{z'p}{2}\sigma$.

The Hamiltonian H_{tot_i} at each monomer is minimized with respect to the choice of amino acid by choosing the smallest of A samples from the distribution of H_{tot_i}—again we wish to estimate the minimum of many samples of a Gaussian. By way of (6) [13], we find that

$$H_{\mathrm{tot}}^{\mathrm{min}} \simeq -\sqrt{2}\, N \sigma_{\mathrm{tot}_i} \sqrt{\ln A} \,. \tag{12}$$

When the trained sequence is in one of the p target structures, the average energy of the sequence is given by

$$E_{\mu}^{\mathrm{min}} \simeq \frac{H_{\mathrm{tot}}^{\mathrm{min}}}{p} \simeq -\sqrt{\frac{z'}{p}}\, N \sigma \sqrt{\ln A} \,. \tag{13}$$

Equation (13) and results from simulation are plotted in Fig. 1 for $p = 1$. Apart from a prefactor of 0.847, the predicted dependence of well depth on A is in good agreement with observation. Calculations for $p > 1$ are ongoing and will be presented elsewhere.

Comparing the minimum copolymer energy (7) and the minimum energy of the trained sequence (13) yields

$$p_{\mathrm{max}} \simeq \frac{\ln A}{\ln \kappa} \,. \tag{14}$$

Capacity from information theory.—The capacity of a protein may also be derived via information theory. Consider the transmission of a message, which has been encoded as an N letter sequence. The message is decoded empirically by constructing the protein corresponding to the sequence (either *in vitro* or via computer simulation), allowing it to fold and observing the p most occupied, and consequently lowest, target conformations.

The information retrieved by learning a single conformation may be determined as follows. Given κ^N possible compact conformations, the information contained in one conformation is equivalent to the number of bits necessary to express a number between 1 and κ^N, viz., $\ln_2(\kappa^N)$. Since the p target configurations are assumed to be independent, the total retrieved information scales linearly with p; that is, $I_R = pN \ln_2 \kappa$.

The information transmitted may be similarly determined. Since the number of sequences grows as A^N, the information associated with a sequence is $\ln_2(A^N)$, and the total transmitted information [14] is $I_T = N \ln_2 A$.

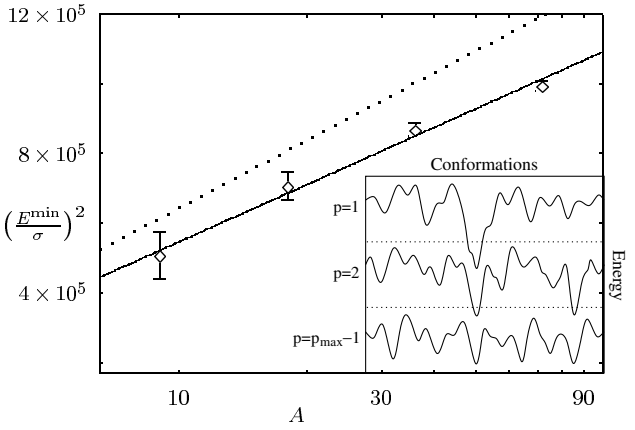

FIG. 1. Square of protein stability $\frac{E^{\mathrm{min}}}{\sigma}$ as a function of the number of amino acid species A (log-linear). Proteins were trained to fold to a single $6 \times 6 \times 6$ conformation with periodic boundary conditions by optimization over sequence space under constant composition. The dotted line was generated by $(13)|_{p=1}$; introducing the prefactor 0.847 gives the solid line. Data are shown for $A = 9$, 18, 36, and 72 species, for each of which the mean and standard deviation were calculated from 12 runs with independent random pair potentials. Inset: energy landscapes of sequences trained to be thermodynamically stable in a one, two, and $p_{\mathrm{max}} - 1$ target conformations. As the number of targets increases, the depth to which the target wells can be trained diminishes. At $p = p_{\mathrm{max}}$, the wells are lost among nearby fluctuations.

Information theory dictates that the information retrieved must not be greater than the information transmitted, that is,

$$pN \ln_2 \kappa \leq N \ln_2 A \,. \tag{15}$$

It readily follows that the bound on p is

$$p_{\mathrm{max}} = \frac{\ln A}{\ln \kappa} \,, \tag{16}$$

which is identical to the result (14) deduced from fluctuations in the energy landscape.

Discussion of capacity.—Our bound on capacity has been derived in two ways: by comparison of the trained and copolymer minimum energies, which depends on the method of training (in our case the superposition rule), and by an information theoretic argument, which does not. The equality of the two results suggests that our constant capacity result is not a shortcoming of the superposition rule.

That our bound on memory is independent of chain length N may seem surprising given that the capacity of a fully connected ANN grows linearly with the number of neurons n. The resolution is that, in both cases, the number of patterns which can be stored is of order the number of connections divided by the number of nodes. In the case of a protein the number of active connections (contacts) is restricted to of order N, whereas for an ANN all n^2 connections are allowed to contribute significantly. The

divisor arises because the amount of information in a pattern is proportional to N and n, respectively.

What happens to the protein energy landscape upon introducing further target conformations? Consider an energy landscape in which there lies a single well of maximal depth. As a second (and, by assumption, independent) well is introduced, the depth of the first is reduced (Fig. 1 inset). As p approaches $p_{\max}$, the typical well depth diminishes such that, at $p = p_{\max}$, the minima are indistinguishable from nearby fluctuations.

For a uniform composition (i.e., a homopolymer), zero conformations are encodable, as expected. Frequently studied binary models allow at most one configuration to be stored, while for a 20 amino acid set, $p_{\max} \simeq 4.7$. In all cases, as p approaches $p_{\max}$, the minima become increasingly nominal. It may be possible to find a binary (e.g., H-P) sequence with a global minimum above an arbitrary compact target. But there is typically of order one sequence per conformation, and the sequence is statistically unlikely to be stable. In this sense, binary models are not accurate representations of proteins.

Application to heteropolymer design and prions.—Our results may be considered in the more general context of heteropolymer engineering and rational drug design. The ability to remember multiple conformations admits a potentially dramatic increase in the variety of heteropolymer function. We have provided arguments that training to superimposed contact maps provides a viable method of designing multiply conforming sequences. To what extent can we exercise control over their occupied conformations?

Abkevich *et al.* [9] observed in simulation what they refer to as kinetic partitioning: some sequences designed to be stable in two conformations initially fold to one structure before later folding to the other. On time scales short by comparison, the distribution over conformations occurs according to kinetic accessibility rather than conformational stability. We are investigating the extent to which temperature can be used to effect a change of the dominant occupied conformation before the onset of equilibrium.

A naturally occurring and much studied heteropolymer thought to possess multiple stable conformations is prion protein [15]. Prions are infectious, transmissible pathogens composed exclusively of the modified protein PrP^{Sc} [5]. The chemical (primary) structure of PrP^{Sc} is identical to the normal prion protein PrP^{C}, but its conformation (tertiary structure) is significantly different. Prion diseases, such as bovine spongiform encephalopathy, Creutzfeldt-Jakob disease, and scrapie of sheep, are believed to result from the conformational conversion of PrP^{C} to PrP^{Sc} and the resulting accumulation of the abnormal protein [5].

Our calculations support the view that prion disease is caused by misfolding to a second stable conformation. Far from being confined to particular or correlated structures, the ability of a protein to take on multiple biologically active conformations is ubiquitous. In addition to pathological proteins such as prions, we conjecture the existence of proteins which fold to multiple biologically *useful* conformations. Definitive observations to this end would have significant implications on our understanding of protein function.

*Present address: Laboratoire de Physique Statistique, Ecole Normale Supérieure, 75231 Paris Cedex 05, France. Electronic address: fink@lps.ens.fr; http://www.tcm.phy.cam.ac.uk/~tmf20/

†Present address: Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom. Electronic address: r.c.ball@warwick.ac.uk

[1] Ken A. Dill and Sun Chan, Nat. Struct. Biol. **4**, 10 (1997).

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

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

[4] D. J. Amit, *Modeling Brain Function* (Cambridge University Press, Cambridge, U.K., 1989).

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

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

[7] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **92**, 1282 (1995).

[8] This assumes that the number of patterns stored does not scale more quickly than N.

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

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

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

[12] Bringing the $\frac{1}{2}$ from (11) into the sum over bonds index bound accounts for frustration.

[13] This estimation is consistent with our use of the central limit theorem provided $\ln A < \frac{z'p}{4}$.

[14] The $A \times A$ pair potential U is part of the decoding apparatus and need not be transmitted.

[15] It has been suggested that the stability of PrP^{Sc} may depend not only on its tertiary (intrachain) structure but also on its quaternary (interchain) structure [M. P. Morrissey and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **96**, 11 293 (1999)]. We do not consider this here.