# Kinetic Capacity of a Protein

Robin C. Ball[§] and Thomas M. A. Fink[†]
*Theory of Condensed Matter, Cavendish Laboratory, Cambridge CB3 0HE, England*
[§]r.c.ball@warwick.ac.uk [†]tmf20@cam.ac.uk [†]http://www.tcm.phy.cam.ac.uk/~tmf20/



The ability of a protein to recognise multiple independent target conformations was demonstrated in [1]. Here we consider the recognition of correlated configurations, which we apply to funnel design for a single conformation. The maximum basin of attraction, as parametrised in our model, depends on the number of amino acid species as $\ln A$, independent of protein length. 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 clarifies the scope and limits of protein and heteropolymer function.

87.14.Ee 36.20.Ey 87.15.Aa 87.15He

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 [2]. These conditions are known as thermodynamic stability and kinetic accessibility, respectively. 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.

The first satisfactory method of protein design, introduced by Shakhnovich in 1994 [3], relies on the correlation between stability and accessibility: stable sequences are found to fold more quickly as well. Minimising the energy or relative energy over sequence space while the conformation remains quenched to the target [3,4] yields protein sequences which fold much more rapidly than random heteropolymers of equal length. We have provided evidence, nonetheless, that the most stable sequences are not the fastest folding, and that a reduction in stability allows significant gain in efficiency [5].

In this Letter we investigate the introduction of a folding funnel above the target conformation in the protein energy landscape [6]. Our method of design relies on the technique of training to multiple targets discussed in [1]. Unlike the independent conformations previously considered, here our patterns are correlated to a single target conformation. We find that the extent of the optimal folding funnel, as parameterised by our model, is smaller than the conformational space and depends on the number of amino acid species available. The influence of alphabet size on the folding performance of *untrained* sequences is considered in [7].

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 unrealistic 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 training to the ideal folding funnel cannot be achieved. Remarkably, the bound on funnel size (in terms of a relaxation length scale) is identical to the thermodynamic capacity derived in [1]. 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 log of the number of amino acid species.

**Generalisation to Weighted Training** In a separate Letter [1] we investigated the design of multi-stable proteins by training to a uniform superposition of contact maps. The typical well depth of a protein of length $N$ embedded in one of the target conformations was found to be

$$E_\mu^{\min} \simeq -\sqrt{\tfrac{z'}{p}} N\sigma\sqrt{\ln A}. \tag{1}$$

where $A$ is the number of amino acid species, $\sigma$ is the standard deviation of the interaction energies and $z' = z - 2$ is the effective coordination number, *i.e.*, the maximum number of local contacts excluding the backbone. After training to a weighted superposition of contact maps, 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,

$$C_{\mathrm{tot}_{ij}} = \sum_{\mu=1}^{p} w_\mu C_{\mu_{ij}}, \tag{2}$$

where $w_\mu$ is the weight associated with conformation $\Gamma_\mu$. The minimum Hamiltonian associated with the total weighted contact map is

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

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

By analogy with calculations in [1], we re-express (3) as a sum over $H_{{\rm tot}_i}$, each minimised by the choice of $S_i$,

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

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

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

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

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

Proceeding as in [1], we approximate the distribution of $H_{{\rm tot}_i}$ by its central limit form; it is a gaussian with variance $\sigma_{{\rm tot}_i}^2 = \frac{z'}{2}\sigma^2\sum_{\mu=1}^{p} w_\mu^2$. This estimation is valid out to $|H_{{\rm tot}_i}| \sim \frac{z'}{2}\sigma\sum_{\mu=1}^{p} w_\mu$.

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

$$H_{{\rm tot}_i} = w_\mu E_{\mu_i} + \sum_{\nu=1,\nu\neq\mu}^{p} w_\nu E_{\nu_i} = H_{\mu_i} + H_{{\rm oth}_i}. \quad (7)$$

Since $H_{\mu_i}$ and $H_{{\rm oth}_i}$ are independently gaussianly distributed with variances

$$\sigma_{\mu_i}^2 = \frac{z'\sigma^2}{2}w_\mu^2 \quad \text{and} \quad \sigma_{{\rm oth}_i}^2 = \frac{z'\sigma^2}{2}\sum_{\nu=1,\nu\neq\mu}^{p} w_\nu^2, \quad (8)$$

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

$$f(H_{\mu_i}|H_{{\rm tot}_i}^{\rm min}) \simeq c\exp\Big(-\frac{\sigma_{{\rm tot}_i}^2}{2\sigma_{\mu_i}^2\sigma_{{\rm oth}_i}^2}(H_{\mu_i} - \frac{\sigma_{\mu_i}^2}{\sigma_{{\rm tot}_i}^2}H_{{\rm tot}_i}^{\rm min})^2\Big), \quad (9)$$

where $c$ is a normalising constant and $\sigma_{{\rm tot}_i}^2 = \sigma_{\mu_i}^2 + \sigma_{{\rm oth}_i}^2$. The value of $H_{\mu_i}$ of maximum likelihood from (9) is given by

$$H_{\mu_i}^{\rm min} = \frac{\sigma_{\mu_i}^2}{\sigma_{{\rm tot}_i}^2}H_{{\rm tot}_i}^{\rm min}, \quad (10)$$

which reduces to

$$H_{\mu_i}^{\rm min} = w_\mu E_{\mu_i}^{\rm min} = \frac{w_\mu^2}{\sum_{\mu=1}^{p} w_\mu^2}H_{{\rm tot}_i}^{\rm min}. \quad (11)$$

The minimum local Hamiltonian corresponds to the smallest of $A$ samples from the distribution of $H_{{\rm tot}_i}$. We approximate the minimum of $A$ samples from a gaussian of zero mean and standard deviation $\sigma_{{\rm tot}_i}$ by [1]

$$H_{{\rm tot}_i}^{\rm min} \simeq -\sqrt{2}\sigma_{{\rm tot}_i}\sqrt{\ln A}. \quad (12)$$

Substituting (12) into (11) and summing over $i$ yields

$$E_\mu^{\rm min} \simeq -\sqrt{z'}N\sigma\sqrt{\ln A}\frac{w_\mu}{\left(\sum_{\mu=1}^{p} w_\mu^2\right)^{\frac{1}{2}}}, \quad (13)$$

This establishes how the minimised Hamiltonian distributes over the individual weighted configurations; for the special case of equal weights it duly reduces to (1).

**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 non-interacting chains with simple kink-jump dynamics, a subsection of $g$ monomer units has relaxation time $\tau(g)$ proportional to $g^2$ [8]. 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 which we do not address here include spatially localised 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 $z'$ non-zero entries in each row and column, $z'N$ non-zero entries in total.

For the state unfolded up to length scale $g$, the protein may be thought of as a chain of $\frac{N}{g}$ blobs, folded to its coarse grained original conformation. Accordingly, the contact map $C(g)$ has $\frac{N}{g}$ intra-blob blocks along the diagonal and $\frac{z'N}{g}$ inter-blob blocks corresponding to nearest neighbour blobs (not along the backbone). Scaling theories for polymer configurations with excluded volume would imply that the average total number of contacts between two neighbouring 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 $\frac{1}{g^2}$.

The total number of conformations (compact or otherwise) available to a protein grows as $\sim \tilde{\kappa}^N$ [8] (not to be confused with $\kappa \simeq 1.85$ [9] for compact structures only); this becomes $\tilde{\kappa}^{\frac{N}{g}}$ for a chain of $\frac{N}{g}$ 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 $\tilde{\kappa}^{(N-\frac{N}{g})}$ configurations. It follows that the entropy gained in folding from a denatured configuration down to a conformation relaxed to length scale $g$ is

$$S(g) = -k_B \frac{N}{g} \ln \tilde{\kappa}. \tag{14}$$

**Training to a Funnel** While an energy minimum significantly below the minimum copolymer energy ensures thermodynamic stability of the target conformation, rapid convergence necessitates 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 $w(g)$,

$$C_{\mathrm{tot}_{ij}} = \sum_{\ln g=1}^{\ln N} w(g) C_{ij}(g). \tag{15}$$

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

$$H_{\mathrm{tot}}^{\min} = \frac{1}{2} \sum_{ij=1}^{N} \sum_{\ln g=1}^{\ln N} w(g) C_{ij}(g) \tilde{U}_{ij}^{*}, \tag{16}$$

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

$$H_{\mathrm{tot}_i}^{\min} = \sum_{\ln g=1}^{\ln N} H_i^{\min}(g), \tag{17}$$

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

$$\sigma^2_{g_i} \simeq \frac{z'g}{2} \Big( \frac{w(g)}{g^2} \Big)^2 \sigma^2, \tag{18}$$

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

$$\sigma^2_{\mathrm{tot}_i} \simeq \sum_{\ln g=1}^{\ln N} \sigma^2_{g_i} \simeq \frac{z'\sigma^2}{2} \int_e^N \frac{dg}{g} g \frac{w^2(g)}{g^4}. \tag{19}$$

Again we wish to establish how the minimised Hamiltonian distributes over weighted configurations unfolded to length scale $g$. Applying the general result (10) yields

$$H_i^{\min}(g) \simeq w(g) E_i^{\min}(g) \simeq \frac{\sigma^2_{g_i}}{\sigma^2_{\mathrm{tot}_i}} H_{\mathrm{tot}_i}^{\min}. \tag{20}$$

Substituting (12) and (18) into (20) and summing over $i$, the minimum energy associated with matching the conformation at scale $g$ can then be estimated as

$$E^{\min}(g) \simeq -\frac{z'}{\sqrt{2}} N \sigma^2 \sqrt{\ln A} \frac{w(g)}{\sigma_{\mathrm{tot}_i} g^3}. \tag{21}$$

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 $T\times$ (14) and (21) gives

$$w(g) \simeq -\frac{\sqrt{2} k_B T \ln \tilde{\kappa}\ \sigma_{\mathrm{tot}_i}}{z'\sigma^2 \sqrt{\ln A}} g^2, \tag{22}$$

and thus $w(g) \propto g^2$. Unfortunately this form for $w$ is inconsistent with a convergent ($N$-independent) evaluation of $\sigma_{\mathrm{tot}_i}$ in (19). 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 (22) into (19) and reducing the domain of integration yields

$$\sigma^2_{\mathrm{tot}_i} \simeq \frac{(k_B T)^2 \ln^2 \tilde{\kappa}}{z'\sigma^2 \ln A} \sigma^2_{\mathrm{tot}_i} \int_e^{g_{\max}} dg, \tag{23}$$

from which it follows that

$$g_{\max} \simeq \frac{z'\sigma^2 \ln A}{(k_B T)^2 \ln^2 \tilde{\kappa}}. \tag{24}$$

The width of our funnel, as parametrised 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 loss in entropy resulting from collapse will be equivalent to $-(14)$ 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 $T_{\mathrm{cp}}$. Equating the minimum copolymer energy $E_{\mathrm{cp}}^{\min}$ from (7) in [1] and $T_{\mathrm{cp}}$ times the loss in entropy $-(14)|_{g=1}$ leads to

$$k_B T_{\mathrm{cp}} \simeq \sigma \frac{\sqrt{z' \ln \kappa}}{\ln \tilde{\kappa}}, \tag{25}$$

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

$$g_{\max} \simeq \frac{\ln A}{\ln \kappa}, \tag{26}$$

which is identical to the form of $p_{\max}$ derived in [1].

**Discussion of Capacity** That the bound on the folding funnel $g_{\max}$ is less than $N$ implies the extent of the achievable folding funnel 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 characterised by a funnel. Above $g_{\max}$, the protein must organise itself into the desired (coarse grained) conformation without the help of kinetic guidance, that is, it must traverse an effective copolymer landscape (Figure 1). What happens to the protein energy landscape upon increasing the width of the funnel? As $g \to g_{\max}$, 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 (Figure 2); 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 realised 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 $g_{\max}$.

**Manipulation of the Energy Landscape** In both the thermodynamic [1] and kinetic contexts, the extent to which the protein energy landscape can be manipulated is limited by $\frac{\ln A}{\ln \kappa}$, where $A$ is the number of amino acid species and $\kappa$ 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 counter-balanced 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.

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 [5]). Accordingly, thermodynamically oriented sequence design need not select for the fastest folding proteins and a reduction in stability admits 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 $\kappa$) 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 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 [10] has extended our notion of heteropolymer behaviour beyond familiar protein collapse. We have presented arguments that the energy landscape may, within limits, be tailored to effect function heretofore unobserved. Further discovery of novel protein mechanisms should prove fascinating.

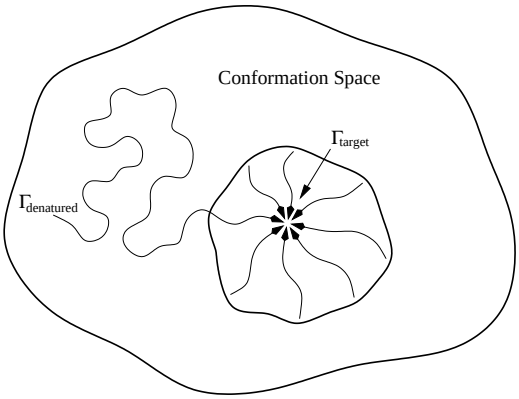


FIG. 1. Folding in the presence of a funnel. The denatured protein wanders through conformation space until it matches the target structure coarse-grained to length scale $g_{\max}$, after which the funnel quickly guides the protein towards the target.

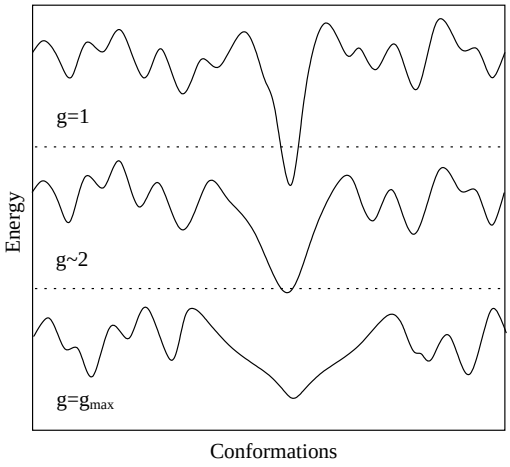


FIG. 2. Energy landscapes of sequences trained to have increasingly broad funnels. Maximising stability (top) corresponds to a deep, narrow well. As the length scale $g$ to which the funnel extends increases, the depth of the target well is reduced; at $g = g_{\max}$, the slope of the funnel is no longer sufficient to provide a free energy minimum (bottom).

[1] Thomas M. A. Fink and Robin C. Ball, submitted to Phys. Rev. Lett. (2000).

[2] Ken A. Dill and Sun Chan, Nature Struct. Biol. **4**, 10 (1997).
[3] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[4] A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, Proc. Natl. Acad. Sci. USA **92**, 1282 (1995).
[5] Thomas M. Fink and Robin C. Ball, Physica D **107**, 199 (1997).
[6] Thomas M. A. Fink, *Inverse Protein Folding, Hierarchical Optimisation and Tie Knots*, Ph.D. thesis, University of Cambridge (1998).
[7] J.-R. Garel, T. Garel and H. Orland, J. Phys. France, **50**, 3067 (1989).
[8] Pierre-Gilles de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, UK, 1979).
[9] Vijay S. Pande *et al.*, J. Phys. A: Math. Gen. **27**, 6231 (1994).
[10] Stanley B. Prusiner, Proc. Natl. Acad. Sci. USA, **95**, 13363 (1998).
[11] Paul M. Harrison *et al.*, J. Mol. Biol. **286**, 593 (1999).