# Gooey tutorial 2: Correlated protein folding landscape

ICFP M2 – Advanced Biophysics

Protein folding involves a rugged landscape with multiple metastable states. As discussed for the random energy model during the lectures, this may cause the dynamics to slow down considerably. This prevents random amino acid sequences from folding reproducibly, while natural ones evade the slow-down thanks to the minimal amount of frustration displayed by their native fold (a "fold" is a specific 3D conformation of the protein).

The picture of complete lack of correlations displayed by the random energy model (sometimes abbreviated as REM) is however an overly pessimistic one. Indeed, it presents the energies of two folds that differ by a single protein contact as completely independent variables, while common sense suggests that they should have similar energies. Here we study a model, the so-called generalized random energy model (GREM), which includes this feature and assigns similar energies to states with a high overlap in the list of their contacts. We first study the entropy of a partially folded chain as a function of overlap in Sec. 1. We then introduce the generalized random energy model in Sec. 2, and analyze the resulting energy correlations as a function of overlap. Such correlations in energy make the folding free energy landscape smoother than in the original random energy model. As a result, a native fold with a very low energy (as prescribed by the minimal frustration principle) generates an energy "funnel" around itself, implying that folding from a neighboring state is somewhat akin to sliding down a free energy slope (Sec. 3). Finally, we propose an optional discussion of the effect of a rugged energy landscape on the coil-globule transition.

## 1 Entropy as a function of overlap

Consider a protein fold of reference with $N$ residues. We additionally consider alternative folds, in which a fraction $q$ of the residues are in the same position as in the reference, and a fraction $1-q$ of the residues deviate from it [Fig. 1(a)]. These deviations are composed of several "excursions" away from the reference fold, and we denote their respective lengths (counted in numbers of residues) as $\ell_1$, $\ell_2$ *etc.*

1.1 Under the coarse approximation (justified in the optional question below) that excursion $j$ is a random walk of length $\ell_j$ that chooses one orientation among $z$ at each step, how many possible realizations are there for an excursion of that length? This approximation boils down to neglecting both self-avoidance and the constraint for the chain to reconnect to the reference fold at the end of the excursion.

1.2 Assuming that the positions and sizes of the excursions are fixed, show that there are $z^{(1-q)N}$ possible realizations of such excursions under the aforementioned approximation.

1.3 Now consider that there are as many choices of excursion positions and sizes as there are choices of $(1-q)N$ detached monomers among $N$. Deduce that to dominant order in $N \to \infty$ the entropy of the chain reads

$$S(q) = Nk_B[(1-q)\ln z - q\ln q - (1-q)\ln(1-q)]. \tag{1}$$

### [Optional] Loop closure and the Poland-Scheraga model

A more sophisticated approach considers the closure constraint of each excursion. Denoting the swelling exponent of the chain by $\nu$, show using a scaling reasoning that the number of possible realizations of an excursion of length $\ell$ under this constraint is proportional to $z^\ell/\ell^{3\nu}$. Now instead of considering the thermodynamic ensemble where the total number of reference monomers $qN$ and of excursion monomers

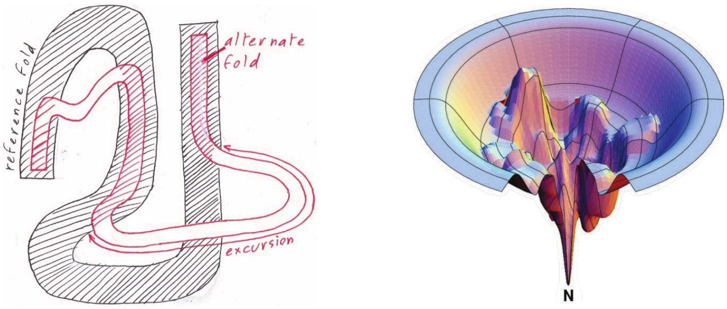

Figure 1: Representations of partially overlapping folds in real space and in energy space. (a) Excursion model representing a reference protein fold in black. A second fold of the same protein is represented as a red ribbon; a fraction $q$ of its residues overlaps with the black fold (hatched red ribbon), and a fraction $1-q$ departs from it (unhatched ribbons). This non-overlapping fraction can be composed of one or several separate "excursions". (b) Energy landscape in the vicinity of a reference fold with low energy (the downward peak), where the horizontal axes are a metaphorical representation of the degrees of freedom of the protein. In this horizontal plane, the radial coordinate is $1-q$, *i.e.*, folds that have a large overlap with the reference fold are close to the center of the plot.

$(1-q)N$ are fixed, introduce the fugacities $r$ and $w$ associated with these numbers. and show that the resulting partition function is $Z(r,w)=\sum_{k=0}^{+\infty}R(r)^{k+1}E(zw)^k$, where

$$R(r)=\sum_{\ell_j=1}^{+\infty}r^{\ell_j}\qquad\text{and}\qquad E(zw)=\sum_{\ell_j=1}^{+\infty}\frac{(zw)^{\ell_j}}{\ell_j^{3\nu}},\tag{2}$$

are the partition functions associated with a regular segment and with an excursion, respectively. In the case where the excursions behave as polymer globules ($\nu=1/3$), compute the values of $r$ and $w$ associated with the limit $N\to+\infty$ with fixed $q$. Deduce that the error committed by using Eq. (1) is negligible in the limit of large $N$.

This model has been widely used to determine the statistics of opening double-stranded DNA. In this context, the regular fractions represent double-standed segments while the excursions stand for so-called DNA bubbles where the two strands separate. The ratio $w/r$ can then be interpreted as a Boltzmann factor $e^{-\beta\epsilon}$, where $\epsilon$ is the DNA pairing energy per base pair. Show that this model has a phase transition towards the opening up of a macroscopic fraction of the DNA strand for finite $\epsilon$ that is second order for $1<3\nu\leqslant 2$ and first order for $3\nu>2$.

## 2 Energy distribution as a function of overlap in the GREM

Now that we are equipped with an expression for the entropy associated with the collection of folds with overlap $q$, we study a model for the distribution of their energies. In this generalized random energy model, we index all possible contact between two residues $j$ and $k$ by an integer $\mu$. Therefore $\mu\equiv\{j,k\}$ and so $\mu\in[1..N(N-1)/2]$. A randomly chosen energy $\epsilon_\mu$ is assigned to each of these possible contacts, and a protein fold is characterized by the list of all $\{\Delta_\mu\}_{\mu\in[1..N(N-1)/2]}$, where $\Delta_\mu=1$ if contact $\mu$ is realized (*i.e.*, residues $j$ and $k$ are in contact in the fold considered), and $\Delta_\mu=0$ otherwise. As a result, the Hamiltonian of the system reads

$$\mathcal{H}(\mathbf{\Delta})=\sum_{\mu=1}^{N(N-1)/2}\epsilon_\mu\Delta_\mu,\tag{3}$$

where $\mathbf{\Delta}$ denotes the set of all $\Delta_\mu$. We only consider globular states of the protein, and assume that each residue interacts with $z$ neighbors in such a configuration, which implies that for all configurations considered there are exactly $Nz/2$ values of $\mu$ for which $\Delta_\mu=1$ (thus $\sum_\mu\Delta_\mu=Nz/2$ for any $\mathbf{\Delta}$).

In the spirit of the random energy model, the values of the contact energies are independent random variables whose distributions are identical and given by

$$\forall\mu \quad P(\epsilon_\mu) = \frac{1}{\sqrt{2\pi b^2}} e^{-\epsilon_\mu^2/2b^2}. \tag{4}$$

2.1 Consider a single protein fold with contacts $\mathbf{\Delta}$. Show that the distribution of the energies of the fold takes the same form as in the random energy model, yielding the density of states

$$\rho_{\mathbf{\Delta}}(E) = \frac{1}{\sqrt{\pi N z b^2}} e^{-E^2/Nzb^2}. \tag{5}$$

To do so you may write $\rho_{\mathbf{\Delta}}(E) = \langle \delta[E - \mathcal{H}(\mathbf{\Delta})] \rangle$, and express the Dirac delta as $\delta(x) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}\lambda}{2\pi} e^{i\lambda x}$.

2.2 Now consider two folds $a$ and $b$ with contacts $\mathbf{\Delta}^a$ and $\mathbf{\Delta}^b$ respectively. We define their overlap $q$ as the fraction of contacts that they have in common, *i.e.*,

$$q = \frac{2}{Nz} \sum_\mu \Delta_\mu^a \Delta_\mu^b. \tag{6}$$

This definition is not identical to the one used in Sec. 1, but we nonetheless combine the results from these two slightly different approaches in Sec. 3. We denote the probability that the energy of fold $a$ is between $E_a$ and $E_a + \mathrm{d}E_a$ *and* that the energy of $b$ is between $E_b$ and $E_b + \mathrm{d}E_b$ as $\rho_{\mathbf{\Delta}^a,\mathbf{\Delta}^b}(E_a, E_b)\, \mathrm{d}E_a\, \mathrm{d}E_b$. Show that the joint density of states for the energies of $a$ and $b$ reads

$$\begin{aligned} \rho_{\mathbf{\Delta}^a,\mathbf{\Delta}^b}(E_a, E_b) &= \langle \delta[E_a - \mathcal{H}(\mathbf{\Delta}^a)]\, \delta[E_b - \mathcal{H}(\mathbf{\Delta}^b)] \rangle \\ &= \frac{1}{\pi N z b^2 \sqrt{1-q^2}} \exp\left\{ -\frac{1}{2Nzb^2} \left[ \frac{(E_a - E_b)^2}{1-q} + \frac{(E_a + E_b)^2}{1+q} \right] \right\} \end{aligned} \tag{7}$$

2.3 Using Bayes' theorem [*i.e.* $P(B|A) = P(A,B)/P(A)$], prove that the conditional density of states is given by

$$\rho_{\mathbf{\Delta}^a,\mathbf{\Delta}^b}(E_b|E_a) = \frac{1}{\sqrt{\pi N z b^2 (1-q^2)}} \exp\left[ -\frac{(E_b - qE_a)^2}{Nzb^2(1-q^2)} \right]. \tag{8}$$

Comment on the $q \to 0$ and the $q \to 1$ limit. Show that the average energy of a fold that has an overlap $q$ with fold $a$ reads

$$\langle E(q) \rangle = qE_a. \tag{9}$$

## [Optional] The REM as the many-body interactions limit of the GREM

The model presented above is actually a special case of the generalized random energy model. The more general version assigns an energy not to a single inter-residue contact, but to a combination of $p$ contacts, where $p \in \mathbb{N}^*$. The Hamiltonian then reads

$$\mathcal{H}(\mathbf{\Delta}) = \sum_{\mu_1 < \mu_2 < \ldots < \mu_p} \epsilon_{\mu_1,\mu_2,\ldots,\mu_p} \Delta_{\mu_1} \Delta_{\mu_2} \ldots \Delta_{\mu_p} \tag{10}$$

In other words, the contribution $\epsilon_{\mu_1,\mu_2,\ldots,\mu_p}$ is added to the energy of the fold if and only if contacts $\mu_1, \mu_2, \ldots, \mu_p$ are all realized at the same time. Such contributions are distributed according to

$$P(\epsilon_{\mu_1,\mu_2,\ldots,\mu_p}) = \sqrt{\frac{(Nz/2)^{p-1}}{2\pi b^2 p!}} \exp\left[ -\left(\epsilon_{\mu_1,\mu_2,\ldots,\mu_p}\right)^2 \frac{(Nz/2)^{p-1}}{2b^2 p!}, \right] \tag{11}$$

which ensures that the energy of the chain remains extensive and that the single-fold density of state is still given by Eq. (5). The case studied previously corresponds to $p = 1$. Compute the joint energy density $\rho_{\mathbf{\Delta}^a,\mathbf{\Delta}^b}(E_a, E_b)$ as a function of $E_a$, $E_b$ and $q$, and show that the random energy model is recovered in the $p \to +\infty$ limit. Use this result to comment on the effect of many-body interactions on the physics of protein folding.

## 3 Funneling and the mean folding landscape

The situation studied in Secs. 1 and 2 is intuitively similar to the picture of Fig. 1(b). To understand this, consider a reference fold with a low energy $E_a$, as is the case for a minimally frustrated native fold. Now consider a fold that has an overlap $q$ with the reference fold. The larger the overlap, the smaller the average energy of the fold [as in Eq. (9)]. There are fluctuations in the energy around this average, and therefore points in Fig. 1 with the same radial coordinate (*i.e.*, the same $q$) can have different energies, as described by Eq. (8). Folds that are close to but different from the reference fold thus tend to slide towards it down the energy slope as if guided by a funnel. In the presence of thermal agitation, the sliding is however hindered by the abundance of states with a small overlap, implying that small overlaps are entropically favored as shown in Eq. (1).

3.1 Write down and plot the average free energy landscape $F(q)$ as a function of the overlap $q$. How low does the energy $E_a$ have to be for the landscape to be tilted towards the reference state?

3.2 Show that if $q$ is a continuum variable, then the reference state is never a local energy minimum. Remembering that a fold has a finite number of contacts, argue that a situation where $F(1-2/Nz) > F(1)$ does funnel the protein all to way towards folding.

3.3 What is the condition on $E_a$ for this situation to occur? Discuss this condition in view of the discussion of minimal frustration in the main lecture.

Note that the glass transition in the generalized random energy model is more complex than in the random energy model: as the temperature is lowered, the configuration space first becomes fragmented into basins within which the configurations have a high overlap. Transitions between basins are frozen out, but transitions within them aren't. As the temperature is lowered further the basins gradually shrink and the systems eventually becomes completely glassy.

## [Optional] When is a molten globule actually glassy?

We now go back to a random energy model with energy distribution

$$\rho(E) = \frac{\Omega}{\sqrt{2\pi z N b^2}} \exp\left[\frac{(E - N\epsilon)^2}{2zNb^2}\right] \qquad \text{with} \qquad \Omega = e^{Ns_g}. \tag{12}$$

In this equation, $\epsilon$ and $s_g$ denote the average energy per residue in the globule state and the entropy per residue in an interaction-less globule. Assuming the energy per residue in the coil state is zero and that the entropy per residue is given by $s_0 = s_g + k_B$, compute the temperature $T_c$ at which the protein transitions from a coil ($T > T_c$) to a globule ($T < T_c$). Using the expression for the glass temperature from the main lecture, plot a phase diagram showing the regions where the protein is a coil, a globule and a glass as a function of temperature as well as its effective "flexibility" $s_0$. Discuss which values of $s_0$ are the most amenable for protein folding.