

### Folding & Unfolding

1. *Designed Random Energy Model (REM)*: Consider a protein model in which for a given sequence and structure, the energy is randomly taken from the Gaussian probability density

$$p(E) = \frac{1}{\sqrt{2\pi\Sigma^2}} \exp\left(-\frac{E^2}{2\Sigma^2}\right).$$

The total number of structures is  $\Omega_{str}$ , while the number of sequences is  $\Omega_{seq} \gg \Omega_{str}$ .

(a) A particular *sequence* has a (unique) native structure of energy  $E_N$ . Calculate and plot the energy  $E(T)$  of this sequence as a function of temperature  $T$ .

(b) For a particular *structure*, we attempt to design a good sequence by Monte Carlo sampling of representative sequences at a ‘temperature’  $\tau$ . Calculate and plot the designed native energies  $E_N(\tau)$  as a function of the design temperature  $\tau$ .

\*\*\*\*\*

2. **(Optional)** *Charged Random Energy Model*: Use the random energy model to investigate the freezing of a charged heteropolymer. Assume that there are  $g^N$  possible globular states of the polymer, whose energies are randomly selected from a Gaussian distribution of mean zero, and variance

$$\sigma^2 = u^2 N + c^2 \left(\frac{Q^2}{R}\right)^2.$$

The second term in the above formula is a rough estimate of the variations in Coulomb energy from different ways of distributing a charge  $Q$  over a volume of size  $R$ .

(a) Find the energy  $E_c$  at which the entropy vanishes, and the corresponding freezing temperature  $T_c$ .

(b) For compact globular states, how should  $Q^2$  scale with  $N$  for the freezing temperature to be asymptotically independent of  $N$ ?

\*\*\*\*\*

3. *Amino-acid interactions*: What can we learn by combining the Random Energy Model with commonly used interaction potentials between amino acids?

(a) Find a  $20 \times 20$  matrix of interactions  $U(a, a')$  amongst amino acids, and calculate the mean  $\langle U \rangle$  and variance  $\langle U^2 \rangle_c$  of its elements. The commonly used Miyazawa–Jernigan (MJ) interaction matrix can be found in S. Miyazawa and R.L. Jernigen, *J. Mol. Biol.* **256**, 623 (1996). (Table 3 of this publication is available on the web-page for assignments.)

(b) Model the possible configurations of a protein by the ensemble of compact self-avoiding walks on a cubic lattice. (All lattice sites are visited by compact walks.) Calculate the number  $n$  of non-polymeric nearest neighbor interactions for such configurations on an  $N = L \times L \times L$  lattice, and deduce the ratio  $n/N$  for large  $N$ .

(c) The number of compact walks on a cubic lattice asymptotically grows as  $g^N$ , with  $g \approx 1.85$ . Use this in conjunction with the results from parts (a) and (b) to estimate the folding temperature  $T_c$  of a random sequence of amino-acids, and the corresponding energy  $E_c$ .

**(Optional)** (d) Select a protein, find its amino-acid sequence and construct a contact matrix corresponding to its structure. Use the interaction matrix from part (a) to estimate the energy of the native structure, and calculate the ratio  $E_N/E_c$ .

\*\*\*\*\*

**4. Kinetics of protein folding:** [Adapted from Gutin *et al.*, J. Chem. Phys. **108**, 6466 (1998).] Assume protein folding proceeds through a folding nucleus which has the free energy  $F^\ddagger = E^\ddagger - k_B T \log M^\ddagger$ . The folding nucleus serves as a transition state for the folding reaction. The typical folding time needed to climb over this free energy barrier is

$$t = \tau_0 \exp\left(\frac{F^\ddagger - F}{k_B T}\right),$$

where  $T$  is the temperature, and  $\tau_0$  is an elementary time step.

(a) Use a random energy model to calculate  $F$  as a function of temperature  $T$ , and calculate the folding time  $t(T)$  for two regimes  $T > T_c$  and  $T < T_c$ . Plot  $\ln t(T)$  as a function of  $1/T$ .

(b) Consider a limit of  $T \rightarrow \infty$  and express the folding time as a function of the total number of conformations  $M = g^N$  and the number of states in the folding nucleus  $M^\ddagger$ . Interpret your result.

(c) Find a temperature  $T_{opt}$ , which provides the fastest folding, compare it to  $T_c$ . Compare the optimal folding time with the folding time from “non-designed” REM at  $T_c$ . Make conclusions about folding kinetics for random sequences (REM) and designed sequences (designed REM).

\*\*\*\*\*

**5. Pulling polymers:** Using optical tweezers, it is now possible to pull on the two ends of a single molecule. (Actually the tweezers pull on latex balls that are attached to the ends of the polymer; a complication that we shall ignore.) In the presence of the force  $\vec{F}$  pulling on the ends of the polymer, there is an additional energy term

$$\delta E = -\vec{F} \cdot \vec{R},$$

where  $\vec{R} = \vec{r}_N - \vec{r}_1$  is the end-to-end distance (between the first and  $N^{\text{th}}$  monomers) of the chain.

(a) For an ideal polymer, the number of configurations with an end-to-end distance of  $\vec{R}$  is given by the usual Gaussian formula

$$\Omega_N(\vec{R}) = \frac{g^N}{(2\pi N a^2/3)^{3/2}} \exp\left(-\frac{3R^2}{2N a^2}\right).$$

By integrating the Boltzmann weight over all  $\vec{R}$ , calculate the (Gibbs) partition function  $Z(N, F, T)$  at a temperature  $T$ . Using this result, obtain the mean extension  $R_F = k_B T \partial \ln Z / \partial F$  along the direction of the force  $\vec{F}$ .

(b) For other cases in which  $\Omega_N$  does not have a simple form (such as for self-avoiding polymers), it is still possible to obtain the *linear response* of the polymer to small force. To this end, expand the Boltzmann weight  $\exp(\vec{F} \cdot \vec{R} / k_B T)$  to second order in  $\vec{F}$ , and hence show that

$$R_F = \frac{1}{3k_B T} \langle R^2 \rangle_0 F + \mathcal{O}(F^3),$$

where  $\langle R^2 \rangle_0$  is the mean end-to-end squared distance of the polymer in the absence of the force.

(c) Dimensional analysis suggests that quite generally the extension-force curve for polymers should have the form

$$\frac{R_F}{\sqrt{\langle R^2 \rangle_0}} = \Phi \left( \frac{F \sqrt{\langle R^2 \rangle_0}}{k_B T} \right).$$

The left hand side is a dimensionless extension; on the right hand side a dimensionless combination involving the force appears as the argument of an unknown function  $\Phi$ . At large forces  $F$ , the polymer becomes stretched such that  $R_f \propto N$ . For self-avoiding polymers  $\sqrt{\langle R^2 \rangle_0} \approx a N^\nu$  with  $\nu \approx 0.59$ . Use these facts to deduce a non-linear behavior  $R_F \propto F^\lambda$  for the extension at large force, and give the value of the exponent  $\lambda$ .

\*\*\*\*\*

**6. Denaturing DNA by force:** Obtain the phase diagram of DNA pulled by a force  $\vec{F}$ , by generalizing the Poland-Scheraga model as follows:

(a) By integrating over the position vectors, show that the (Gibbs) partition function of DNA of length  $N$  can be decomposed into products of contributions from double-stranded rods and single stranded bubbles, as

$$Z(N, F) = \sum_{\ell_1, \ell_2, \ell_3, \dots} R(\ell_1) B(\ell_2) R(\ell_3) \dots, \quad \text{with} \quad \ell_1 + \ell_2 + \ell_3 + \dots = N.$$

(b) Treat the double stranded segments as rigid rods of fixed length  $a\ell$ . By integrating over all orientations in three dimensions show that

$$R(\ell) = w^\ell \times \frac{\sinh(\beta F a \ell)}{\beta F a \ell},$$

where  $w = e^{-\beta \varepsilon}$ , and  $\varepsilon$  is the energy gain of forming the double strand.

(c) Treat the double stranded loop as two random walks of length  $\ell$  connected at the two end points. Integrating over all separations of the two end points show that

$$B(\ell) = \frac{s}{\ell^{3/2}} \left[ g^2 \exp \left( \frac{\beta^2 F^2 a^2}{12} \right) \right]^\ell.$$

- (d) Examine the problem in a (grand canonical) ensemble with variable DNA lengths  $N$ , additionally weighted by a factor of  $z^N$ . Give the expressions for the (Laplace) transformed  $\tilde{B}(z)$  and  $\tilde{R}(z)$  in this ensemble in terms of the (Bose) sums  $f_m^+(x) = \sum_{\ell=1}^{\infty} x^\ell / \ell^m$ .
- (e) Show that the strands become fully separated at a critical point satisfying  $\tilde{R} = \tilde{B}^{-1} = (s\zeta_{3/2})^{-1}$ , where  $\zeta_{3/2} \equiv f_{3/2}^+(1) \approx 2.612$ .
- (f) For  $s = 1$ , plot the phase diagram of the model in the coordinates  $(w/g^2)$  and  $(\beta Fa)$ .

\*\*\*\*\*

**7. Denaturing RNA by force:** By pulling on the ends of RNA, the hydrogen bonds can be broken to yield a stretched polymer. Let us model the partially denatured state as a sequence of linear segments with no hydrogen bonds and ‘blobs’ which are hydrogen bonded (opposite to the case of DNA). Assume that the force carrying backbone of the molecule is made up of the linear segments, and that the RNA blobs carry no force (similar to the loop in problem 2). After integrating over the position vectors, the (Gibbs) partition function of an RNA of length  $N$  can be written as

$$Z(N, F) = \sum_{\ell_1, \ell_2, \ell_3, \dots} P(\ell_1)R(\ell_2)P(\ell_3)\dots, \quad \text{with} \quad \ell_1 + \ell_2 + \ell_3 + \dots = N.$$

The contributions of linear and blob segments are respectively

$$P(\ell) = g^\ell \exp\left(\frac{F^2 a^2 \ell}{6k_B^2 T^2}\right), \quad \text{and} \quad R(\ell) = f^\ell \frac{A}{\ell^{3/2}},$$

where  $f$  and  $g$  are constant entropic factors.

- (a) Exploit the mathematical similarity to the Poland–Scheraga model to evaluate the grand partition function of the model.
- (b) Identify the force  $F_c$  at which denaturation starts.
- (c) Sketch the fraction of denatured sites as a function of force, clearly indicating the nature of the singularity at  $F_c$ .

\*\*\*\*\*

**8. (Optional) Pulling RNA:** The server at <http://bioserv.mps.ohio-state.edu/rna/> gives force extension curves for RNA based on secondary structure calculations. Use this server to examine force extension curves for: (a) a uniform sequence; (b) an alternating sequence of G and C; (c) an alternating sequence of A and U; and (d) an actual RNA sequence. (Choose sequences of roughly the same length.) Comment on the general characteristics of these curves. Does any of them resemble the theoretical result from the previous problem?

\*\*\*\*\*