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 Ω_{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.

• Average number of states with energy E is $n(E) = \Omega_{\rm str} p(E)$. In microcanonical distribution entropy is defines as:

$$S(E) = k_B \ln n(E) \approx k_B \ln \Omega_{\rm str} - k_B \frac{E^2}{2\Sigma^2},$$

where we have neglected the small term $\ln\sqrt{2\pi\Sigma^2}$. At large temperatures we determine the connection between equilibrium energy and temperature:

$$\frac{1}{k_B T} = \frac{\partial(S/k_B)}{\partial E} = -\frac{E}{\Sigma^2} \qquad \rightarrow \qquad E(T) = -\frac{\Sigma^2}{k_B T}.$$

This is valid only for $T > T_f$ where protein can fold and unfold fast. At $T < T_f$ protein is trapped in native state once it folds in it. Temperature of transition is obtained from 'tangent construction' as discussed in class:

$$\frac{1}{k_B T_f} = \frac{\partial(S/k_B)}{\partial E} = -\frac{E(T_f)}{\Sigma^2} = \frac{S\big(E(T_f)\big)/k_B}{E(T_f) - E_N} = \frac{\ln \Omega_{\rm str} - E(T_f)^2/2\Sigma^2}{E(T_f) - E_N}.$$

After some algebraic manipulation this equation translates to quadratic equation:

$$0 = \beta_f^2 - 2\beta_f\beta_N + \beta_C^2,$$

where we have introduced $\beta_N = -E_N/\Sigma^2$ and $\beta_C^2 = 2\ln\Omega_{\rm str}/\Sigma^2$. Freezing temperature $T_f$is higher than T_C, where the entropy vanishes. We must pick the solution of the quadratic equation $\beta_f = \beta_N - \sqrt{\beta_N^2 - \beta_C^2}$, which satisfies condition $\beta_f < \beta_C < \beta_N$. Freezing temperature is thus:

$$T_f = \frac{\Sigma^2/k_B}{-E_N - \sqrt{E_N^2 - 2\Sigma^2 \ln \Omega_{\rm str}}}$$

At temperature T_f we have first order transition where equilibrium energy jumps from $E(T_f) = E_N + \sqrt{E_N^2 - 2\Sigma^2 \ln \Omega_{\rm str}}$ to E_N. Energy $E(T)$ of sequence as a function of energy is plotted in Fig. 1.

$$E(T) = \begin{cases} -\frac{\Sigma^2}{k_B T}, & T > T_f \\ E_N, & T < T_f \end{cases}$$

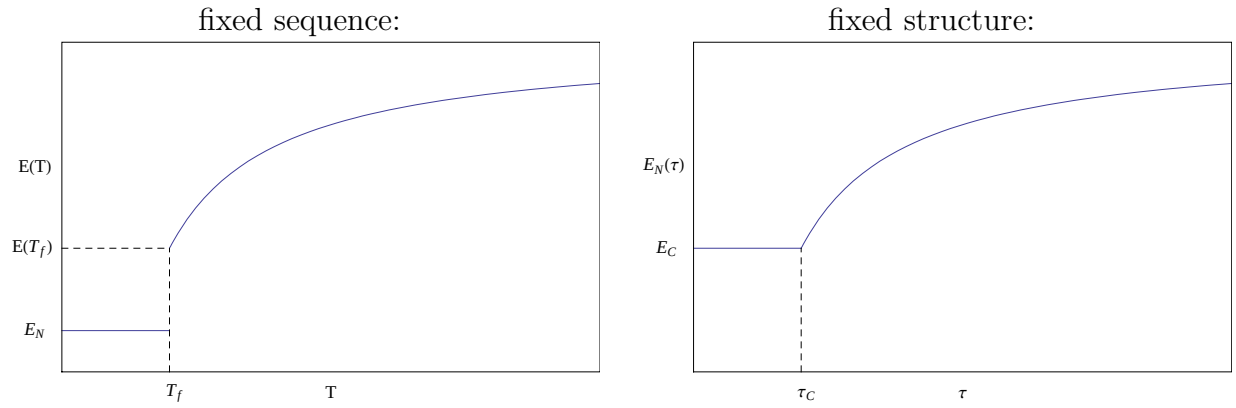


Figure 1: Schematic plots of energy dependence $E(T)$ for fixed sequence and $E_N(\tau)$ for fixed structure. For fixed sequence there is first order transition at freezing temperature T_f. For fixed structure there is second order transition at temperature τ_C.

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

• Average number of states with energy E is $n(E) = \Omega_{\rm seq} p(E)$. In microcanonical distribution entropy is defined as:

$$S(E) = k_B \ln n(E) \approx k_B \ln \Omega_{\rm seq} - k_B \frac{E^2}{2\Sigma^2},$$

where we have neglected the small term $\ln \sqrt{2\pi\Sigma^2}$. At large temperatures we determine the connection between equilibrium energy and temperature:

$$\frac{1}{k_B T} = \frac{\partial (S/k_B)}{\partial E} = -\frac{E}{\Sigma^2} \qquad \rightarrow \qquad E(T) = -\frac{\Sigma^2}{k_B T}.$$

When we will make Monte Carlo sampling at a 'temperature' τ, we will observe energy $E(\tau)$ defined above with probability ≈ 1. This is true only when energy $E > E_C$, where $S(E_C) = 0$ with corresponding temperature τ_C.

$$S(E_C) = 0 = k_B \ln \Omega_{\rm seq} - k_B \frac{E_C^2}{2\Sigma^2} \quad \rightarrow \quad E_C = -\sqrt{2\Sigma^2 \ln \Omega_{\rm seq}} \ \text{ and } \ k_B \tau_C = \frac{\Sigma}{\sqrt{2 \ln \Omega_{\rm seq}}}$$

For temperature below τ_C continuous picture is not good any more. In general there are a few discrete energy states below E_C. When we make Monte Carlo sampling we expect to observe energy near E_C, where we have much more states.

$$E_N(\tau) = \begin{cases} -\frac{\Sigma^2}{k_B \tau}, & \tau > \tau_C \\ E_C, & \tau < \tau_C \end{cases}$$

*****

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.

- Entropy is defined like in problem 1:

$$\frac{S(E)}{k_B} = N \ln g - \frac{E^2}{2\sigma^2} = N \ln g - \frac{E^2}{2\left(u^2 N + c^2 (Q^2/R)^2\right)}$$

Energy E_c at which the entropy vanishes is

$$E_c = -\sqrt{2\sigma^2 N \ln g} = -\sqrt{2N\left(u^2 N + c^2 (Q^2/R)^2\right) \ln g}$$

and the corresponding freezing temperature T_c is

$$k_B T_c = -\frac{\sigma^2}{E_c} = -\sqrt{\frac{u^2 N + c^2 (Q^2/R)^2}{2N \ln g}}.$$

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

- For the freezing temperature to be asymptotically independent of N, Q^2/R should scale as $\sqrt{N}$. In the compact globular state in 3 dimensions R scales as $N^{1/3}$. This means that Q should scale as:

$$Q \sim R^{1/2} N^{1/4} \sim N^{5/12}.$$

*****

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 20×20 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.)

- The Miyayava-Jernigan interaction matrix, where each element is in RT units:

	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly
Cys	-5.44	-4.99	-5.8	-5.5	-5.83	-4.96	-4.95	-4.16	-3.57	-3.16
Met	-4.99	-5.46	-6.56	-6.02	-6.41	-5.32	-5.55	-4.91	-3.94	-3.39
Phe	-5.8	-6.56	-7.26	-6.84	-7.28	-6.29	-6.16	-5.66	-4.81	-4.13
Ile	-5.5	-6.02	-6.84	-6.54	-7.04	-6.05	-5.78	-5.25	-4.58	-3.78
Leu	-5.83	-6.41	-7.28	-7.04	-7.37	-6.48	-6.14	-5.67	-4.91	-4.16
Val	-4.96	-5.32	-6.29	-6.05	-6.48	-5.52	-5.18	-4.62	-4.04	-3.38
Trp	-4.95	-5.55	-6.16	-5.78	-6.14	-5.18	-5.06	-4.66	-3.82	-3.42
Tyr	-4.16	-4.91	-5.66	-5.25	-5.67	-4.62	-4.66	-4.17	-3.36	-3.01
Ala	-3.57	-3.94	-4.81	-4.58	-4.91	-4.04	-3.82	-3.36	-2.72	-2.31
Gly	-3.16	-3.39	-4.13	-3.78	-4.16	-3.38	-3.42	-3.01	-2.31	-2.24
Thr	-3.11	-3.51	-4.28	-4.03	-4.34	-3.46	-3.22	-3.01	-2.32	-2.08
Ser	-2.86	-3.03	-4.02	-3.52	-3.92	-3.05	-2.99	-2.78	-2.01	-1.82
Asn	-2.59	-2.95	-3.75	-3.24	-3.74	-2.83	-3.07	-2.76	-1.84	-1.74
Gln	-2.85	-3.3	-4.1	-3.67	-4.04	-3.07	-3.11	-2.97	-1.89	-1.66
Asp	-2.41	-2.57	-3.48	-3.17	-3.4	-2.48	-2.84	-2.76	-1.7	-1.59
Glu	-2.27	-2.89	-3.56	-3.27	-3.59	-2.67	-2.99	-2.79	-1.51	-1.22
His	-3.6	-3.98	-4.77	-4.14	-4.54	-3.58	-3.98	-3.52	-2.41	-2.15
Arg	-2.57	-3.12	-3.98	-3.63	-4.03	-3.07	-3.41	-3.16	-1.83	-1.72
Lys	-1.95	-2.48	-3.36	-3.01	-3.37	-2.49	-2.69	-2.6	-1.31	-1.15
Pro	-3.07	-3.45	-4.25	-3.76	-4.2	-3.32	-3.73	-3.19	-2.03	-1.87

	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro
Cys	-3.11	-2.86	-2.59	-2.85	-2.41	-2.27	-3.6	-2.57	-1.95	-3.07
Met	-3.51	-3.03	-2.95	-3.3	-2.57	-2.89	-3.98	-3.12	-2.48	-3.45
Phe	-4.28	-4.02	-3.75	-4.1	-3.48	-3.56	-4.77	-3.98	-3.36	-4.25
Ile	-4.03	-3.52	-3.24	-3.67	-3.17	-3.27	-4.14	-3.63	-3.01	-3.76
Leu	-4.34	-3.92	-3.74	-4.04	-3.4	-3.59	-4.54	-4.03	-3.37	-4.2
Val	-3.46	-3.05	-2.83	-3.07	-2.48	-2.67	-3.58	-3.07	-2.49	-3.32
Trp	-3.22	-2.99	-3.07	-3.11	-2.84	-2.99	-3.98	-3.41	-2.69	-3.73
Tyr	-3.01	-2.78	-2.76	-2.97	-2.76	-2.79	-3.52	-3.16	-2.6	-3.19
Ala	-2.32	-2.01	-1.84	-1.89	-1.7	-1.51	-2.41	-1.83	-1.31	-2.03
Gly	-2.08	-1.82	-1.74	-1.66	-1.59	-1.22	-2.15	-1.72	-1.15	-1.87
Thr	-2.12	-1.96	-1.88	-1.9	-1.8	-1.74	-2.42	-1.9	-1.31	-1.9
Ser	-1.96	-1.67	-1.58	-1.49	-1.63	-1.48	-2.11	-1.62	-1.05	-1.57
Asn	-1.88	-1.58	-1.68	-1.71	-1.68	-1.51	-2.08	-1.64	-1.21	-1.53
Gln	-1.9	-1.49	-1.71	-1.54	-1.46	-1.42	-1.98	-1.8	-1.29	-1.73
Asp	-1.8	-1.63	-1.68	-1.46	-1.21	-1.02	-2.32	-2.29	-1.68	-1.33
Glu	-1.74	-1.48	-1.51	-1.42	-1.02	-0.91	-2.15	-2.27	-1.8	-1.26
His	-2.42	-2.11	-2.08	-1.98	-2.32	-2.15	-3.05	-2.16	-1.35	-2.25
Arg	-1.9	-1.62	-1.64	-1.8	-2.29	-2.27	-2.16	-1.55	-0.59	-1.7
Lys	-1.31	-1.05	-1.21	-1.29	-1.68	-1.8	-1.35	-0.59	-0.12	-0.97
Pro	-1.9	-1.57	-1.53	-1.73	-1.33	-1.26	-2.25	-1.7	-0.97	-1.75

The average contact energy for this interaction matrix is $\langle U \rangle = -3.17RT$. Variance of matrix elements is $\langle U^2 \rangle_c = 2.17R^2T^2$. This are the energies per kilo mole if we are interested in energy per contact we should divide by Avogadro number and note that $k_B = R/N_A$. Per contact we have $\langle U \rangle = -3.17k_BT$ and $\langle U^2 \rangle_c = 2.17k_B^2T^2$.

(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.

- In cubic lattice each lattice point has 6 nearest neighbors. In compact walks all points are occupied by protein. 2 nearest neighbors are neighbor base pairs along the protein, thus for each lattice point we have 4 contacts, except for the points on the surface, but in large N limit we can ignore them. Thus the number of nearest neighbor interactions approaches $n/N \to 4/2 = 2$ for large N. We need to divide by two because every contact is counted twice.

(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.

- Average energy and the variance of the protein energy are:

$$\begin{aligned} \langle E \rangle &= n \langle U \rangle = 2N \langle U \rangle \\ \sigma^2 &= n \left\langle U^2 \right\rangle_c = 2N \left\langle U^2 \right\rangle_c \end{aligned}$$

As in first two problems we define entropy:

$$\begin{aligned} \frac{S(E)}{k_B} &= N \ln g - \frac{(E - \langle E \rangle)^2}{2\sigma^2}, \\ \frac{1}{k_B T_c} &= \frac{\partial S(E)/k_B}{E} = -\frac{E - \langle E \rangle}{\sigma^2}. \end{aligned}$$

To determine the numerical values we will assume that k_BT is evaluated at room temperature with $T = 300^\circ K$. Energy of random sequence of amino-acids is defined by $S(E_c) = 0$:

$$\frac{E_c}{N} = \frac{\langle E \rangle}{N} - \sqrt{\frac{2\sigma^2 \ln g}{N}} = 2 \langle U \rangle - \sqrt{4 \langle U^2 \rangle_c \ln g} \approx -6.33k_BT - 2.31k_BT = -8.64k_BT = -0.22eV$$

Folding temperature T_c:

$$T_c = -\frac{\sigma^2}{k_B(E_c - \langle E \rangle)} = \frac{2 \times 2.17k_B^2T^2}{2.31k_B^2T} = 1.87 \times T = 560^\circ K$$

At room temperature proteins are already folded.

(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.

• We will use proteins from problem 4 and corresponding amino acid locations are found in PDB files. Energy of the native state is calculated as

$$E_N = \sum_{i,j} \Delta^N_{i,j} U(a_i, a_j),$$

where $\Delta^N_{i,j}$ is contact map. This element is 1 if two amino acids are closer than R_c and 0 otherwise. In provided article we find the estimate for $R_c \approx 6.5$Å, where we are checking distance between C^α atoms that are marked as `CA` in PDB files. Note that we are checking the connectivity only for indices where $|i-j| \geq 3$ as discussed in class. From the provided PDB files we can find energy of the native state and the number of contacts n.

file	N	E_N/N	E_N/E_c	n/N	E_N/n
1TEN.PDB	89	$-5.10\ k_BT$	0.59	1.70	$-3.01\ k_BT$
2EBN.PDB	285	$-6.03\ k_BT$	0.70	1.82	$-3.32\ k_BT$
4HHB_A.PDB	141	$-5.53\ k_BT$	0.64	1.70	$-3.25\ k_BT$

We expected that E_N would be bigger than E_C. It looks that in REM we have overestimated number of contacts, because we have observed $n/N \approx 1.7 - 1.8$, but on the other hand in our proteins $N \approx 90 - 290$ was not really large, thus the amino acids on the surface that reduce the n/N ratio are important. For a cubic box $N = L \times L \times L$ we can actually find analytic expression for number of contacts:

$$2n = 8 \times 1 + 12(L-2) \times 2 + 6(L-2)^2 \times 3 + (L-2)^3 \times 4,$$

where this terms mean that 8 endpoints have only one neighbor, points on the side of the cube has two neighbors, points on the face of the cube have tree neighbors and bulk points have four neighbors. In the limit of large L we obtain the same result $n/L^3 \to 2$. In our proteins $L = \sqrt[3]{N} \approx 4.5 - 6.6$ is extremely small. For such small lengths L we can doubt in the correctness of the random walk method on cubic lattice point. From corrected formula for n we find better estimate for number of contacts for compact walk on cubic lattice and thus better estimate for E_c:

file	N	n/N	E_c/N	E_N/E_c
1TEN.PDB	89	1.33	$-6.09\ k_BT$	0.84
2EBN.PDB	285	1.54	$-6.92\ k_BT$	0.87
4HHB_A.PDB	141	1.42	$-6.46\ k_BT$	0.86

The ratio between E_N and E_c is still smaller than 1. Another thing we notice is that for the shortest sequence average contact energy E_N/n was surprisingly smaller than the average of all elements in interaction matrix. For the other two proteins average contact energy was bigger but not that much as we would expected.

*****

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 τ_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 $1/T$.

• Consider a random energy model is consisting of M states whose energies are randomly distributed according to a Gaussian probability of mean E_0 and variance Σ. The partition function at temperatures $T > T_c$ is given by

$$Z(\beta < \beta_c) = M\langle e^{-\beta E}\rangle = M \exp\left(-\beta E_0 + \frac{\beta^2\Sigma^2}{2}\right). \tag{1}$$

Here, the random energy model is characterized by a total of M states whose energies are randomly distributed according to a Gaussian probability of mean E_0 and variance Σ. At low temperatures the statistical weight is dominated by a single state of energy E_c, and hence

$$Z(\beta > \beta_c) = e^{-\beta E_c}. \tag{2}$$

Matching Z and $E = -\partial Z/\partial\beta$ from the above two expressions leads to the usual results of $\beta_c = \sqrt{2\ln M}/\Sigma$ and $E_c = E_0 - \Sigma\sqrt{2\ln M}$.

Note that the folding time has the form $t = \tau_0 Z/Z^\ddagger$, with $Z^\ddagger \equiv M^\ddagger e^{-\beta E^\ddagger}$ the 'partition function' of the folding nucleus. Thus

$$\ln\frac{t(\beta < \beta_c)}{\tau_0} = \ln\frac{M}{M^\ddagger} - \beta(E_0 - E^\ddagger) + \frac{1}{2}\beta^2\Sigma^2, \tag{3}$$

$$\ln\frac{t(\beta > \beta_c)}{\tau_0} = \ln\frac{1}{M^\ddagger} + \beta(E^\ddagger - E_c), \tag{4}$$

corresponding to a quadratic portion at small β and a linear segment at large β as depicted in Fig. 2. The order of energies in this figure is $E_c < E^\ddagger < E_0$.

(b) Consider a limit of $T \to \infty$ 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.

• In the limit of high temperature, i.e. for $\beta \to 0$, the energy exponents are irrelevant, and

$$\frac{t}{\tau_0} = \frac{M}{M^\ddagger}, \tag{5}$$

which is simply the inverse probability of randomly finding a nucleus state amongst all possible states. The proposed folding time simply generalizes this, replacing $M/M^\ddagger$ with $Z/Z^\ddagger$ at finite temperatures.

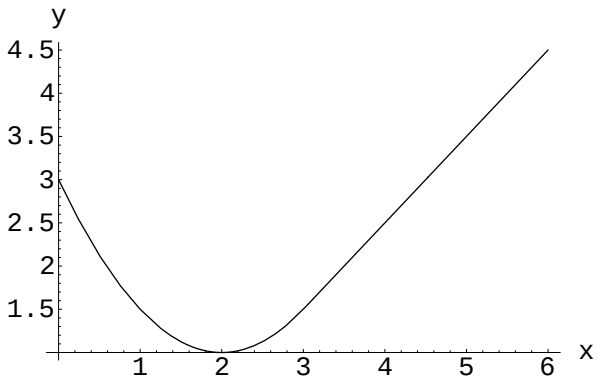


Figure 2: $y = \ln t(\beta)$ as a function of $x = \beta = 1/(k_B T)$, in arbitrary units, with the folding temperature at $\beta_c = 3$.

(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).

- Figure 2 indicates a minimum time in folding, whose location is obtained as

$$\frac{\partial \ln[t(\beta < \beta_c)/\tau_0]}{\partial \beta} = -(E_0 - E^\ddagger) + \beta_F \Sigma^2 = 0\,, \quad \Longrightarrow \quad \beta_F = \frac{E_0 - E^\ddagger}{\Sigma^2}\,. \tag{6}$$

The ratio of fast-folding temperature T_F to freezing temperature T_c is given by

$$\frac{T_F}{T_c} = \frac{\beta_c}{\beta_F} = \frac{\sqrt{2\ln M}}{\Sigma} \cdot \frac{\Sigma^2}{E_0 - E^\ddagger} = \frac{E_0 - E_c}{E_0 - E^\ddagger}\,. \tag{7}$$

The optimal folding time is reduced from the counting result at infinite temperature to

$$\ln \frac{t(\beta_F)}{\tau_0} = \ln \frac{M}{M^\ddagger} - \frac{(E_0 - E^\ddagger)^2}{2\Sigma^2}\,. \tag{8}$$

Manipulation of previous expressions yields the folding time at the freezing point β_c as

$$\ln \frac{t(\beta_c)}{\tau_0} = \ln \frac{M^2}{M^\ddagger} - \frac{\sqrt{2\ln M}(E_0 - E^\ddagger)}{\Sigma}\,. \tag{9}$$

Subtracting the above two expressions we find the ratio the two time scales as

$$\ln \frac{t(\beta_c)}{t(\beta_F)} = \ln M + \frac{(E_0 - E^\ddagger)(E^\ddagger - E_c)}{2\Sigma^2}\,. \tag{10}$$

The relation of the above dynamical model with the equilibrium designed REM becomes apparent by noting that both rely on the ratios of partition functions: In the designed REM the partition function Z of the regular REM competes with the weight $Z_N \equiv e^{-\beta E_n}$ of the

native state; in the dynamic model the ratio of Z to $Z^{\ddagger}$ is considered. With the designed state, the equilibrium model has a phase transition as the average energy is reduced to E_f (see lecture notes). From the comparison of partition functions, it is easy to deduce that for $E^{\ddagger} = E_f$, the temperature of optimal folding is precisely T_f, the equilibrium freezing temperature of the designed REM. The relation between the folding nucleus and the native state is thus apparent.

*****

5. *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

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

- It is obvious that partition function can be decomposed in this way:

$$Z(N,F) = \sum_{\ell_1,\ell_2,\ell_3,\cdots} \int d^3\vec{r}_1 \mathcal{R}(\ell_1,\vec{r}_1) \int d^3\vec{r}_2 \mathcal{B}(\ell_2,\vec{r}_2-\vec{r}_1) \int d^3\vec{r}_3 \mathcal{R}(\ell_3,\vec{r}_3-\vec{r}_2)\cdots,$$

where $\vec{r}_i$ represent end point locations of double-stranded rods ($\mathcal{R}$) and single stranded bubbles ($\mathcal{B}$). Each part is almost independent of the next one, they are only connected with the positions where rods and bubbles are joined together. Integrals can be evaluated one by one from right to left and this is equivalent to

$$\begin{aligned} Z(N,F) &= \sum_{\ell_1,\ell_2,\ell_3,\cdots} \int d^3\vec{s}_1 \mathcal{R}(\ell_1,\vec{s}_1) \int d^3\vec{s}_2 \mathcal{B}(\ell_2,\vec{s}_2) \int d^3\vec{s}_3 \mathcal{R}(\ell_3,\vec{s}_3)\cdots \\ &= \sum_{\ell_1,\ell_2,\ell_3,\cdots} R(\ell_1)B(\ell_2)R(\ell_3)\cdots. \end{aligned}$$

(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 ε is the energy gain of forming the double strand.

- Integrating over all possible directions of rod yields

$$R(\ell) = \frac{1}{2}\int_{-1}^{1} d\mu\; e^{-\beta\epsilon\ell} e^{\beta F a\ell\mu} = w^{\ell} \times \frac{\sinh(\beta F a\ell)}{\beta F a\ell}.$$

As usual we have ignored the unimportant factors of 4π.

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

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

- This is similar problem to the problem 1

$$B(\ell) \;=\; \int d^3\vec{R} \left[\frac{g^\ell}{2\pi\sigma^2} \exp\left(-\frac{R^2}{2\sigma^2}\right)\right]^2 \exp(\beta \vec{F}\cdot\vec{R}),$$

where $\sigma^2 = \ell a^2/3$. Introducing $\sigma^2 = 2\tilde{\sigma}^2$, we obtain the same integral as in problem 1

$$\begin{aligned} B(\ell) &= \frac{1}{(8\pi\tilde{\sigma}^2)^{3/2}} \int d^3\vec{R} \frac{(g^2)^\ell}{(2\pi\tilde{\sigma}^2)^{3/2}} \exp\left(-\frac{R^2}{2\tilde{\sigma}^2}\right) \exp\left(\beta\vec{F}\cdot\vec{R}\right) \\ &= \frac{1}{(8\pi\tilde{\sigma}^2)^{3/2}} g^{2\ell} \exp\left(\frac{\tilde{\sigma}^2\beta^2F^2}{2}\right) \\ &= \frac{s}{\ell^{3/2}} \left[g^2 \exp\left(\frac{\beta^2 F^2 a^2}{12}\right)\right]^\ell . \end{aligned}$$

In the last step we have inserted the value of $\tilde{\sigma}^2 = \ell a^2/6$ and combined all unimportant constants in s.

(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 $f_m^+(x) = \sum_{\ell=1}^\infty x^\ell/\ell^m$.

- As in class we introduce Laplace transformation:

$$\begin{aligned} \tilde{B}(z) &= \sum_{\ell=1}^\infty z^\ell B(\ell) = \sum_{\ell=1}^\infty \frac{s}{\ell^{3/2}} \left[z g^2 \exp\left(\frac{\beta^2 F^2 a^2}{12}\right)\right]^\ell = s f^+_{3/2}\left(z g^2 \exp\left(\frac{\beta^2 F^2 a^2}{12}\right)\right), \\ \tilde{R}(z) &= \sum_{\ell=1}^\infty z^\ell R(\ell) = \sum_{\ell=1}^\infty \frac{z^\ell w^\ell}{\beta F a \ell} \left[e^{\beta F a \ell} - e^{-\beta F a \ell}\right] = \frac{1}{\beta F a} \ln\left(\frac{1 - z w e^{-\beta F a}}{1 - z w e^{\beta F a}}\right) \end{aligned}$$

Laplace transform of partition function is the sum over all bubbles:

$$\Gamma(z) = \tilde{R}(z) + \tilde{R}(z)\tilde{B}(z)\tilde{R}(z) + \cdots = \frac{1}{\tilde{R}^{-1}(z) - \tilde{B}(z)} .$$

The following discussion is similar to the one in class where we analyzed force-less case. $\tilde{B}(z)$ starts at zero for $z = 0$ and monotonically increases with z until $z^* = 1/\left(g^2 \exp\left(\beta^2F^2a^2/12\right)\right)$, where $B(z^*) = s\zeta_{3/2}$. For $z > z^*$, $\tilde{B}(z)$ diverges. $\tilde{R}^{-1}(z)$ starts at infinity for $z = 0$ and then monotonically decreases to zero at $z = 1/(we^{\beta Fa})$.

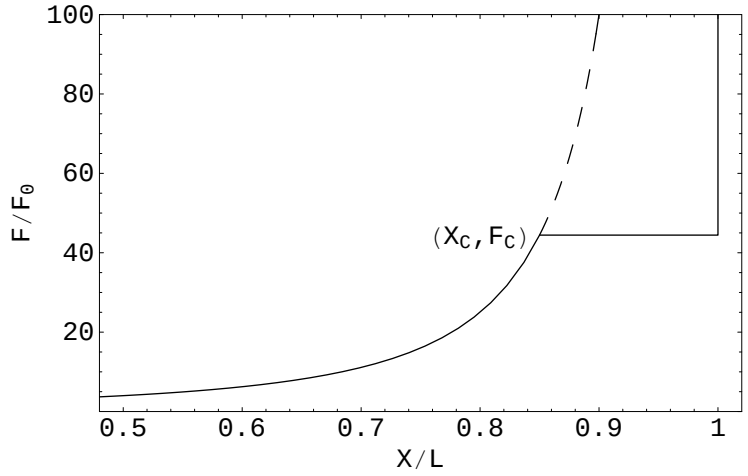
(e) Show that the strands become fully separated at a critical point satisfying $\tilde{R} = \tilde{B}^{-1} = \left(s\zeta_{3/2}\right)^{-1}$, where $\zeta_{3/2} \equiv f^+_{3/2}(1) \approx 2.612$.

• As in class we introduce the average length $\langle L\rangle = z\partial \ln \Gamma/\partial z$ of the DNA and the average number of bound pairs $\langle N_B\rangle = w\partial \ln \Gamma/\partial w$. We are interested in the limit of the very large lengths $\langle L\rangle$, which is equivalent to condition $\tilde{R}^{-1}(z) = \tilde{B}(z)$. The critical point is met, when $\tilde{R}^{-1}(z_c) = \tilde{B}(z_c) = s\zeta_{3/2}$ and $z_c = 1/\left(g^2 \exp\left(\beta^2 F^2 a^2/12\right)\right)$. These conditions also set the critical value of w_c. For $w < w_c$ DNA is fully separated. From expression for $\tilde{R}(z)$ we determine w_c

$$\begin{aligned}
\frac{1 - z_c w_c e^{-\beta F a}}{1 - z_c w_c e^{\beta F a}} &= \exp\left(\frac{\beta F a}{s\zeta_{3/2}}\right) = C, \\
z_c w_c &= \frac{C-1}{C e^{\beta F a} - e^{-\beta F a}} \\
\frac{w_c}{g^2} &= \left\{ \frac{\exp\left(\frac{\beta F a}{s\zeta_{3/2}}\right) - 1}{\exp\left(\frac{\beta F a}{s\zeta_{3/2}}\right) e^{\beta F a} - e^{-\beta F a}} \right\} \exp\left(\beta^2 F^2 a^2/12\right) .
\end{aligned}$$

(f) For $s = 1$, plot the phase diagram of the model in the coordinates (w/g^2) and $(\beta F a)$.

• The phase diagram is plotted below



We see that for large forces, $\beta F a \gg 1$, DNA is fully separated. w_c/g^2 reaches minimum value ≈ 0.045 at $\beta F a \approx 5.71$.

*****

6. *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

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

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.

- As in the Poland-Scheraga model the grand partition function is:

$$\begin{aligned}
\Gamma(z) &= \sum_N z^N Z(N,F) = \frac{1}{\tilde{P}^{-1}(z) - \tilde{R}(z)}, \\
\tilde{P}(z) &= \sum_\ell z^\ell P(\ell) = \frac{zg \exp\left(\frac{F^2 a^2}{6k_B^2 T^2}\right)}{1 - zg \exp\left(\frac{F^2 a^2}{6k_B^2 T^2}\right)}, \\
\tilde{R}(z) &= \sum_\ell z^\ell R(\ell) = A f_{3/2}^+(fz) \ .
\end{aligned}$$

(b) Identify the force F_c at which denaturation starts.

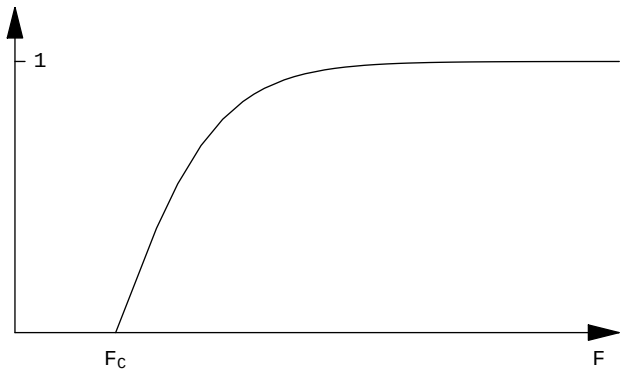
- As in the Poland-Scheraga model, the critical point is located at $z_c = 1/f$ and $\tilde{R}(z_c) = \tilde{P}^{-1}(z_c)$, resulting in

$$\begin{aligned}
A\zeta_{3/2} &= \frac{f}{g \exp\left(\frac{F_c^2 a^2}{6k_B^2 T^2}\right)} - 1, \\
F_c &= \frac{k_B T}{a} \sqrt{6 \ln\left(\frac{f}{g(1 + A\zeta_{3/2})}\right)}.
\end{aligned}$$

This is true only for $f > g(1 + A\zeta_{3/2})$. If this condition is not met, then $F_c = 0$.

(c) Sketch the fraction of denatured sites as a function of force, clearly indicating the nature of the singularity at F_c.

- As in the Poland-Scheraga model denaturated sites (linear segments) start forming for $F > F_c$ and the fraction of denatured sites scales as $(F - F_c)^{\frac{2-c}{c-1}}$, where $c = 3/2$. This means that the fraction of denaturated sites scales linearly with $(F - F_c)$ near F_c. Note that the fact that we have $\exp(F^2 \ldots)$ factor in the $\tilde{P}(z)$ does not affect the scaling behavior, because we are interested in small changes around F_c. Scaling behavior around F_c is completely determined by the scaling behavior of $f_c^+(x) \propto (1 - x)^{c-1}$ near $x = 1$.



*****

7. (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?

- Server will be tested with snR30 sequence with 608 nucleotides found on

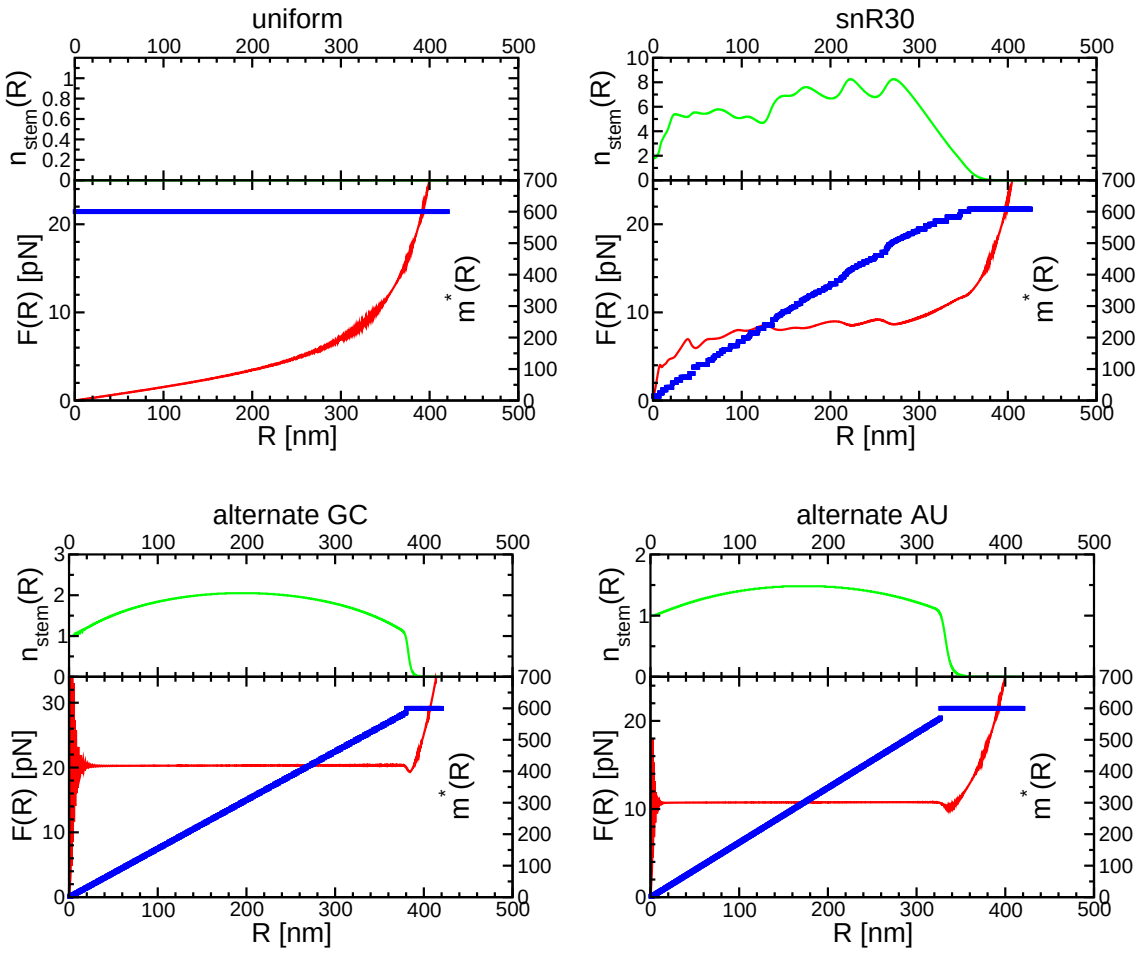
`http://people.biochem.umass.edu/ sfournier/fournierlab/snornadb/snrs/snr30_ta.php`

Sequence:

1	aaccauaguc	ucgugcuagu	ucgguacuau	acagggaagg	gaagucacuc	gcauacgugu
61	gugugcauuu	cuugcuauug	cugcuuagcu	ucucuaaaac	acugggcuac	guuuuucaac
121	gcucgagagg	cagagucuca	aggagccucc	aaugggccuc	acguauucau	cuagauggcg
181	cuucggacaa	cggcaucaca	uaagagaugc	agcuccugac	uucuccucug	aucuucguga
241	ucagaguuuu	gagucgucag	acuacgagca	guuucucuua	gucguugcau	cgggugcugu
301	ugccuuaagc	auguguauau	gggguucggg	ggcuguugcc	augauauaua	uggaugagac
361	agaaguggcc	ccguugacga	guuuaacuua	gauuaaguag	gacgcaugau	cuugagcucu
421	uuuccuauac	uuuguccuau	ggccagcuuu	cuccuuauua	cgaagagauu	gcgggaugug
481	ggugcagagu	gggaaaaucu	gaguucgguc	aucuuuguug	uucguccuac	cgcaguauau
541	uccuaaacac	uaugaaauga	cccuaguugg	uccaugauca	uuuggguaaa	accauacugc
601	agacaucu					

None of the force extension curves on the next page matches particularly well with the theoretical model of the previous problem. The curve for a uniform sequence of 600 elements has $F_c = 0$, and apparently no secondary structure is formed in this case. In the case of an alternating sequence of G and C (A and U) we have to apply certain force $F_c = 20$pN ($F_c = 10$pN) to break the G–C (A–U) primary bonds. The dominant structure is one big line segment with G–C (A–U) primary bonds and one big blob. At critical force G–C (A–U) primary bonds starts braking at the contact of line segment and blob. Eventually all primary bonds are broken and after that extension starts linearly with increasing force $F > F_c$ as in the previous problem. For the real RNA sequence we still see some plateau in the force-extension diagram for forces ≈ 8pN when we start breaking G–C and A–U bonds. The last

bond is broken at force $F_c \approx 14$pN and then we are again in the regime predicted in previous problem.



*****