Protein Interactions

1. *Weight matrices:* You are given a set of binding sites for the E. Coli purine repressor, $PurR$ (see file purR_sites.txt on the assignment page).

(a) Build a weight matrix for $w_i(b)$ for base b at position i for $PurR$. Calculate the information content of the set.

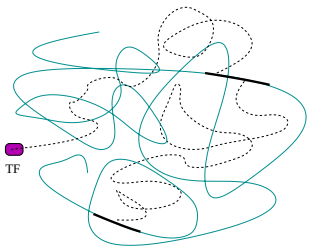
(b) Write a program that feeds random DNA sequences into the weight matrix, and construct a histogram for the resulting weights. Use this histogram to compute the probability distribution of specific binding energies, assuming an effective evolutionary "temperature" equal to the ambient temperature $T^* = 1/(k_B\lambda) = T$.

(c) Assume a *binding threshold* slightly above the average binding energy of the given set of sites. Find all the sequences in the E. coli genome (included on the Assignment page) having binding energy below this threshold. Have you located all the input sequences? Try to move the threshold. How many false-positives do you find? You can consult the provided gene table of E. coli (gene_table.txt) to find out if the detected sequences have any regulatory function.

(d) Consider a simpler model of protein-DNA recognition, where the consensus sequence of DNA of length L provides the highest affinity (the lowest energy of binding), and each mismatch increases the binding energy by a constant ϵ. Calculate the information content of such motif. What is the probability of finding the consensus site in random DNA? How is the probability related to the information content?

*****

2. *Target site location:* Complex transcription machinery in cells is regulated by a set of protein molecules–*transcription factors* (TFs) whose functions can be described as:
• *Receiving a control signal*- This can be the binding or unbinding of a ligand, resulting in initiation or shutting down of the transcription machinery.
• *Finding a specific site on the DNA and binding to it.*



(a) Suppose the protein has to locate a unique binding site on a genome of length M. It may do so by alternately diffusing in solution, and sliding along the DNA, as depicted in Fig. 1. Given a typical TF diameter of 10nm and cytoplasm dynamic viscosity of approximately

0.1 g $s^{-1}cm^{-1}$, estimate D_{3d} for a TF in cytoplasm. (For 1D sliding, one can assume $D_{1d} \approx 0.1 * D_{3d}$.)

(b) Consider the mechanism of search by sliding and 3D diffusion discussed on the lecture. In addition to theses processes, a protein can make occasional hops, i.e. once it dissociates from DNA, it associates again *at the same place*. Calculate the total search time, assuming that upon dissociation a protein will make a hop with a probability $p \approx 0.9$.

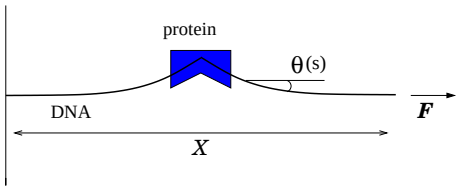
(c) Given the 1D diffusion coefficient D_{1d}, obtain the optimal target location time t_{loc}. The dissociation rate of the proteins from DNA is controlled by the nonspecific binding energy E_{ns}. Estimate E_{ns} for the optimal target location time. Assume $D_{1d} = 1\mu m^2/sec$, $\tau_{3d} = 10^{-3} sec$. Find the location time for $M = 10^6$ base-pairs.

*****

3. *Protein-DNA interaction through bending:* In the worm-like chain (WLC) model, the energy cost of deformed piece of DNA of length L is

$$H = \frac{1}{2}\int_0^L ds\ \frac{\kappa}{R^2(s)},$$

where κ is the bending modulus and $R(s)$ is the local curvature radius. Proteins specifically bound to DNA introduce local "kinks" in the DNA structure. Consider the experimental setup in figure below.



(a) Express the local curvature radius through the local inclination angle $\theta(s)$. Modify the above Hamiltonian to include the applied force F and write it down as $H[\theta(s)]$.

(b) By minimizing H, find the equation for $\theta(s)$. Assuming θ is small, solve the equation and calculate the extension X of the DNA as a function of F. Invert the relation and plot the function $F(X)$.

(c) Calculate the energy cost of the DNA deformation. Given that near the protein, $\theta = 0.5$ and that the energy of specific binding is 20 k_BT, estimate the force at which the protein will "pop" from the DNA. Plot the modified force–extension curve.

*****

4. *Force–extension curve for DNA. II. Entropic elasticity:* Now assume that the DNA is "naked", i.e. there are no proteins attached to it. However, there is still a force applied to its end, and it is not fully extended.

(a) Assume $\theta(s)$ is small. Rewrite the WLC Hamiltonian as a sum over harmonic modes of the DNA "string" θ_q.

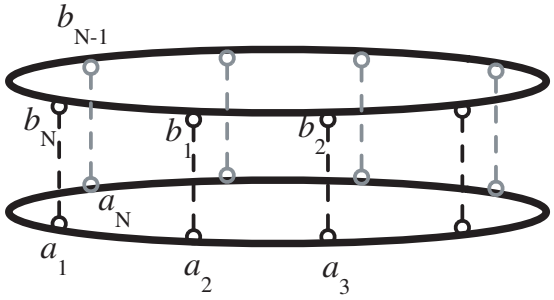
(b) Equipartition requires that each mode will have the average energy of k_BT associated with it. Write the expression for $\langle|\theta_q|^2\rangle$ and calculate $\langle\theta^2(s)\rangle$.

(c) Calculate the extension X of the DNA as a function of force F. Invert the relation and plot the function $F(X)$. Use k_BT and persistence length l_p in your answer.

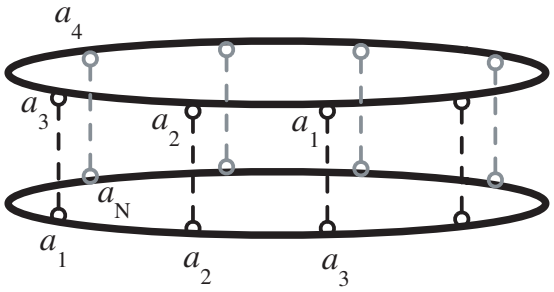
*****

5. *Homodimers versus heterodimers:* In Phys. Rev. Lett. **97**, 178101 (2006), Lukatsky, Zeldovich and Shakhnovich note that proteins are more likely to pair and interact as homodimers (two identical components) than heterodimers (with two distinct parts). They offer a statistical justification for this preference which is partly based on the charecteristics of extreme values. The simplified and analytically tractable model presented in this problem captures this aspect of the explanation.

Heterodimer interface:



Homodimer interface:



We shall assume that the protein binding interfaces are circular rings of exactly N amino-acids. For a given ring, the amino-acids are selected randomly. A heterodimer is constructed by placing two such rings (a and b) in contact, and the resulting binding energy is

$$E_s^{a,b} = \sum_{i=1}^{N} V\left(a_i, b_{i+s}\right) \qquad , \qquad E^{a,b} = \min_s \left\{E_s^{a,b}\right\} .$$

Note that the two rings can be bound after relative shifts by $s = 1, 2, \cdots, N$, and the molecules rotate to achieve the location of minimal energy.

There are two ways to obtain a homodimer: The two sequences can be shifted and matched (not shown in the figure), in which case

$$E^{a,a} = \min_s \{E_s^{a,a}\} \quad , \text{with} \quad E_s^{a,a} = \sum_{i=1}^{N} V\left(a_i, a_{i+s}\right).$$

However, since the rings are at the interface of a larger protein, such matching is generally not possible. The correct arrangement (as in the figure) is to rotate one of the two rings, and then join them, such that

$$E^{a,a_R} = \min_s \{E_s^{a,a_R}\} \quad , \text{with} \quad E_s^{a,a_R} = \sum_{i=1}^{N} V\left(a_{N-i}, a_{i+s}\right).$$

Throughout this problem assume that due to the addition of many pairwise interactions ($N \gg 1$), the energies E_s are Gaussian random variables, and that

$$\langle V(a,b)\rangle = 0, \qquad \text{while} \qquad \langle V(a,b)^2\rangle = \sigma^2.$$

(a) For the heteropolymers, find the mean $\left\langle E^{a,b}\right\rangle$ for $N \gg 1$, and comment on the form of the probability distribution for $E^{a,b}$.

(b) For the un-rotated homodimers, find the probability distribution for $E^{a,a}$, and its mean value. (Hint: Note the number of distinct realization of s.)

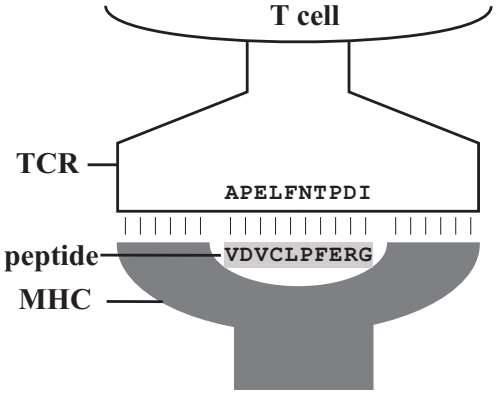
(c) For the rotated homodimers, find the probability distribution for E^{a,a_R}, and its mean value. (Hint: Note the number of distinct interaction terms.)

(d) For randomly selected choices, which ensemble is likely to lead to (i) best binding; (ii) worst binding? If sequences are specifically designed to achieve optimal binding, is there any advantage to homodimers?

*****

6. *Thymic selection of T-cell receptors:* T cells are part of the adaptive immune system; their job is to examine short peptides cut from larger proteins and presented on the surface of cells in blood stream. The strength of binding between a receptor complex on the T-cell, and the peptide presented on another (major histocompatibility) complex, is used to determine whether the peptide comes from a self-protein or is part of a foreign pathogen protein. Pathogens are recognized when the variable T cell receptors (TCRs) bind strongly to foreign peptides; TCRs bind weakly to self-peptides and are thus self-tolerant. To ensure self-tolerance (thereby avoiding auto-immune response), the subset of T cells released to the blood stream is culled from a much larger candidate set in the thymus. In this problem, a simplified model of thymic selection of TCRs is mapped to an extreme value problem.

We shall assume that the relevant binding energy for discriminating between self and foreign peptides is due to an interface of N amino-acids from the peptide, and the TCR.



The starting model thus resembles the previous problem, with

$$E(t,p) = \sum_{i=1}^{N} V(t_i, p_i),$$

where $t \equiv (t_1, t_2, \cdots, t_N)$ and $p \equiv (p_1, p_2, \cdots, p_N)$ indicate the sequences of amino-acids on the TCR and the peptide respectively. A given thymocyte (immature T cell) has some TCR sequence t; it moves around the thymus encountering cells presenting peptides from self-proteins. We shall assume that each thymocyte encounters M such peptides $\{p^{(\alpha)}\}$ for $\alpha = 1, 2, \cdots, M$. It is released into the blood stream only if two conditions are met:

$\star$ It must not bind any self-peptide too strongly. This condition, known as *negative selection* will be modeled by the requirement $E(t, p^{(\alpha)}) > E_n$ for all α (negative energies correspond to stronger binding).

$\star$ It must bind at least one self-peptide moderately. This *positive selection* will be indicated by requiring $E_p > E(t, p^{(\beta)}) > E_n$ for some β.

(a) Show that the above selection criteria are equivalent to $E_n < E_{\rm min}(t) < E_p$, where $E_{\rm min}(t) \equiv \min\{E(t, p^{(\alpha)})\}$ is the strongest binding energy.

(b) For a given TCR amino-acid t_i, let us set

$$\langle V(t_i, p_j)\rangle = \mathcal{E}(t_i), \qquad \text{and} \qquad \langle V(t_i, p_j)^2\rangle - \langle V(t_i, p_j)\rangle^2 = \mathcal{V}(t_i).$$

For large N, what is the probability distribution for $E(t,p)$ (for a given t encountering a random peptide).

(c) What is the probability distribution for $E_{\rm min}(t)$?

(d) Show that for $\ln M \propto N$, the distribution for $E_{\rm min}(t)$ is very narrow, centered around a value proportional to N, while its width does not grow with N.

*****