Protein–DNA 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) **(Optional)** 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.

****

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.*

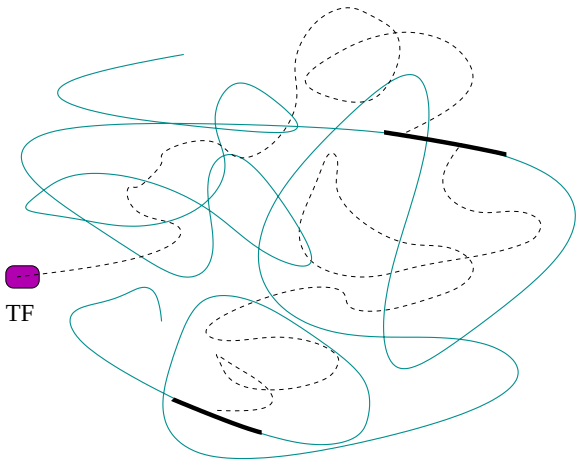

Fig. 1- Schematics of the target location mechanism.
Dashed lines represent 3D diffusion trajectories,
thick black lines are the 1D sliding footprints.

(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) We postulate that the target site is located when the protein traverses the site during one of the 1D sliding events. Assuming that each time the protein slides along the DNA it covers n base-pairs in *completely uncorrelated regions of DNA*, calculate the mean number N of sliding segments needed to locate the target. Assuming that N is large enough, that the motion along the DNA is diffusive, and that the average 3D diffusion time is τ_{3d}, write the expression for the mean target location time $t_{loc}(n)$.

(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. *Force–extension curve for DNA. I. Protein-DNA interaction:* 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 Fig. 2.

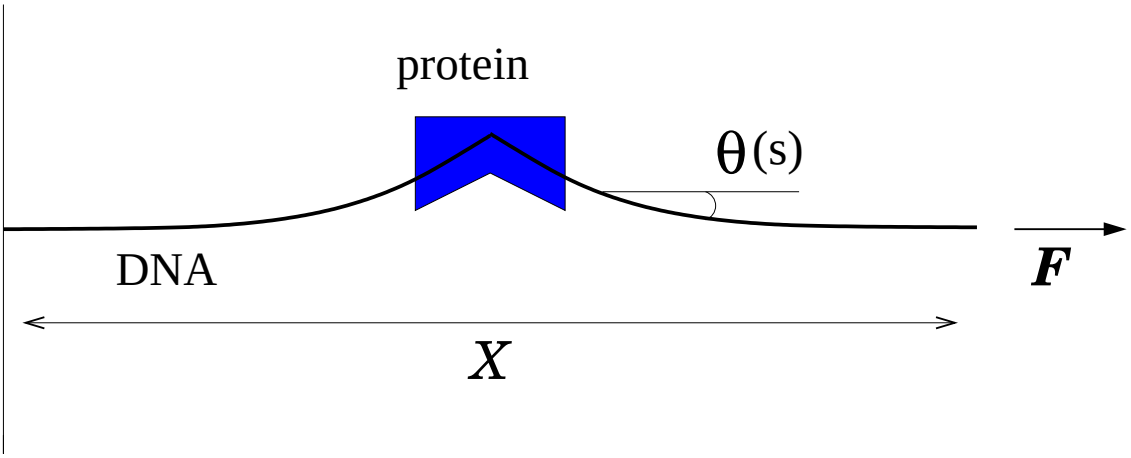

Fig. 2- Force-extension curve measurement setup.

(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.

****