Quantitative principles in biological systems
Problem Set #3
Due by noon on 2026/05/21

1. Information:
    a. The file `hb.csv`[1] contains Hb expression levels for 1,000 points along the anterior-posterior axis across 20 embryos. The data are normalized so that the mean expression has maximum 1 and minimum 0 across positions.
        i. Calculate the mutual information between position and Hb expression, using as rudimentary a method as you can.
        ii. Compare your results to outputs from pre-written packages that you can find online. What factors might underlie differences in the results?
    b. The file `seqs.txt`[1] contains sequences for the promoter region of CRP – which we discussed in Lecture 5 – obtained through random mutations of the native sequence. Each sequence is associated with a batch number representing the expression level of a fluorescent protein under the control of this promoter sequence. The batch B0 represents a sample of all sequences in the experiment.
        i. Calculate and visualize the matrix of the mutual information between the base at each site and the batch number. Interpret your results.
        ii. Generate a random promoter sequence of the same length, assuming that each base is chosen at some probability independent across sites. In which batch would you expect to find this sequence? Try several random sequences. What do your results suggest about the extent of non-specific binding for CRP?

2. Spin glass: Spin glasses can model a wide variety of complex systems. They are a beautiful subject and while they can be arbitrarily complex, they are straightforward and fun to simulate. Let's try to get some intuition on frustration and emergent dynamical behaviors.
    a. Consider a spin glass described by an energy function $E = -\sum_{ij} J_{ij} s_i s_j$, with interactions J_{ij} drawn at random from a normal distribution with zero mean and unit variance. Consider $N \approx 20$ spins, small enough to enumerate every state. Calculate the energy of all states in the system. How many states lie close to the ground state? Are their configurations similar to each other? Repeat your analysis for several instances of randomly drawn interactions and interpret your results relative to hypothetical energy landscapes for protein conformations.
    b. Now consider dynamics on this system given by the simple update rule $s_i \rightarrow \mathrm{sign}\left[\sum_j J_{ij} s_j\right]$, where the sign function returns $+1$ (-1) if the argument is positive (negative).
        i. Convince yourself that the update rule decreases the energy function.
        ii. Simulate the dynamics for $J_{ij} = x_i x_j$, where $\vec{x}$ is an arbitrary binary vector representing some pre-specified configuration. Do the dynamics stop near $\vec{x}$? Repeat your simulation for several instances.
        iii. Now suppose $J_{ij} = J \sum_{\mu=1}^{K} x_i^{(\mu)} x_j^{(\mu)}$, representing K pre-specified configurations. Try $K = 2$ and check whether the dynamics stop near one of the configurations $\vec{x}^{(\mu)}$. Increase K relative to N. Do you find that the stopping points become far away from all $\vec{x}^{(\mu)}$? Interpret your results.
        iv. Take an instance of the above simulation, remove one spin from the system, and repeat your simulation. How do the dynamics compare? What if you remove more spins? Interpret your results.
    c. Suppose that we have enough sequences for a protein family to accurately estimate the frequency $f_i(A_i)$ for the occupancy of residue position i by amino acid type A_i, as well as the two-site frequency $f_{ij}(A_i, A_j)$ for position pair (i, j) by amino acid types (A_i, A_j). What is the maximum entropy model for the distribution $P(A_1, A_2, \ldots, A_N)$ of the joint frequency of all residues?

[1] Bialek's *Biophysics*.