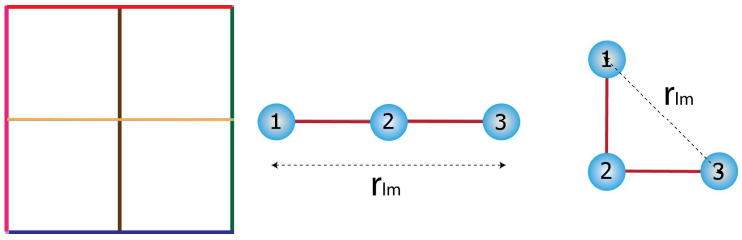
BB 101: Module II

TUTORIAL 4

1. Imagine a protein made of three connected positive charges. The length of the bond between two neighboring proteins is$1\ nm$. This three-charge protein is lying on a 3×3 square lattice in 2D (or a 2D grid connecting 9 lattice sites) as shown below. Color of the grid line denote the spatial inhomogeneity such that all possible conformations/microstates become unique and are not related by rotational symmetry



The Coulomb energy of the protein, in a conformation/microstate i is given by the typical formula for energy,

$$U_i = \sum_{l=1}^{2} \sum_{m=l+1}^{3} \frac{A}{r_{lm}}$$

Where r_{lm} is the distance between charges l and m. Assume$A = 1\ k_B T\ nm$. Note that the charges can only lie on the sites of the lattice and the bonds on the edges.

(a) What is the energy of the protein in the conformation/microstate when all the three charges on a straight line?

(b) What is the energy of the protein in the conformation/microstate that is bent (non-straight; when one bond is making 90^o angle with the other one)?

(c) How many straight conformations are possible on this square lattice?

(d) How many bent conformations are possible on this square lattice?

(e) What is the probability that you will find the protein in a straight structural state or straight macrostate?

(f) What is the probability that you will find the protein in a bent structural state/macrostate?

2. Gene expression is governed by the interaction of DNA with proteins. Imagine that there is a long DNA that has many protein binding sites and you are given multiple copies of a protein that can bind to given DNA. Protein binding can take place only when there is an empty binding site on the DNA and protein can't bind to a site which is already occupied by another protein. You mix DNA and protein. Now, as a function of time you observe that proteins are binding and dissociating from the binding site on the DNA. Let k^+ be the binding rate of proteins (binding rate=number of proteins bound per second). Let k^- be the dissociation (unbinding) rate (number of proteins dissociated per second). Let $\rho(t)$ be the mean/average density of bound proteins at a given time t. (density = number of proteins bound divided by total number of binding sites. Alternatively, it is the fraction of binding sites occupied). Keeping this picture mind, answer the following questions:

(a) Write down the simplest equation for the mean density change ($\frac{\partial \rho}{\partial t}$), if one considers only binding events. Imagine that, at $t = 0$, you have a DNA with no proteins bound and the proteins start binding with a rate k^+ and assume that there is no dissociation. *Hint:* You must address the physical reality that once all the binding sites are fully occupied with proteins, then no more protein can bind. Can you solve the resulting equation and plot mean density as a function of time?

(b) Now imagine that at $t = 0$ all the binding sites are bound with proteins; now, if one considers only dissociation of proteins (no binding), how would the mean density change? Write down an equation for mean density change ($\frac{\partial \rho}{\partial t}$) assuming the dissociation rate as k^-. Can you solve the resulting equation and plot mean density as a function of time? *Hint:* You must address the physical reality that unbinding can't happened if there are no bound proteins

(c) Combine the above two differential equations such that you have an equation for the mean density change ($\frac{\partial \rho}{\partial t}$) when there are both binding and dissociation (consider both k^+ and k^-)

(d) After sometime, the system will reach an "steady-state" where the mean density is a constant. At "steady-state" what will be the mean density?

(e) Now, consider the following limiting cases after you got equation for mean density from part **(d)**

(i) binding and unbinding rates are equal

(ii) binding rate is twice of unbinding rate

(iii) unbinding rate is twice of binding rate

(iv) binding rate is zero

(v) unbinding rate is zero

3. Claude Shannon is known as the “father" of information theory. While working in Bell Labs in 1948, he wrote a famous paper titled “A mathematical theory of communication" in which he derived a formula for “entropy" that would become world famous as “Shannon entropy". He showed that entropy is

$$S = -K \sum_i \; p_i \ln p_i$$

Substitute $p_i = \frac{1}{Z} e^{-\frac{U_i}{k_B T}}$ and $K = k_B$ in the above expression and simplify.

Take $-k_B T \ln Z = G$. After simplification, you should get a well-known relation.

4. During evolution, some genes get mutated and the resulting proteins get altered. In biology, it is very useful (and often important) to find out the DNA sequence that is “conserved" during evolution. Entropy can be a simple measure of this conservation (or the lack of it) during evolution. Let us imagine you got 10 DNA sequences (say, from 10 different organism). Each of these sequences have 3 bases as shown below.

AAT

AGT

ATA

ACG

ATT

AGT

ACT

AAC

ATT

AGT

(i) Calculate the entropy (disorder) at each position (column) using following relation

$$S = -k_B \sum_{i=1}^{M} p_i \ln p_i$$

where M is the number of different letters in each position (column) and $p_i = n_i$/N, where n_i is the number of letters of type i in the column, and N is the total number of letters in that position (column).

(ii) Calculating entropy for each position (column)? Find out which position is more “conserved" over evolution and which position is least conserved over evolution

Notes: Those highly conserved positions are likely to have some crucial role in the function/folding of the protein. This also tells you how to use information theory {theory used for communication by electrical engineers {to understand information content in biological sequences.