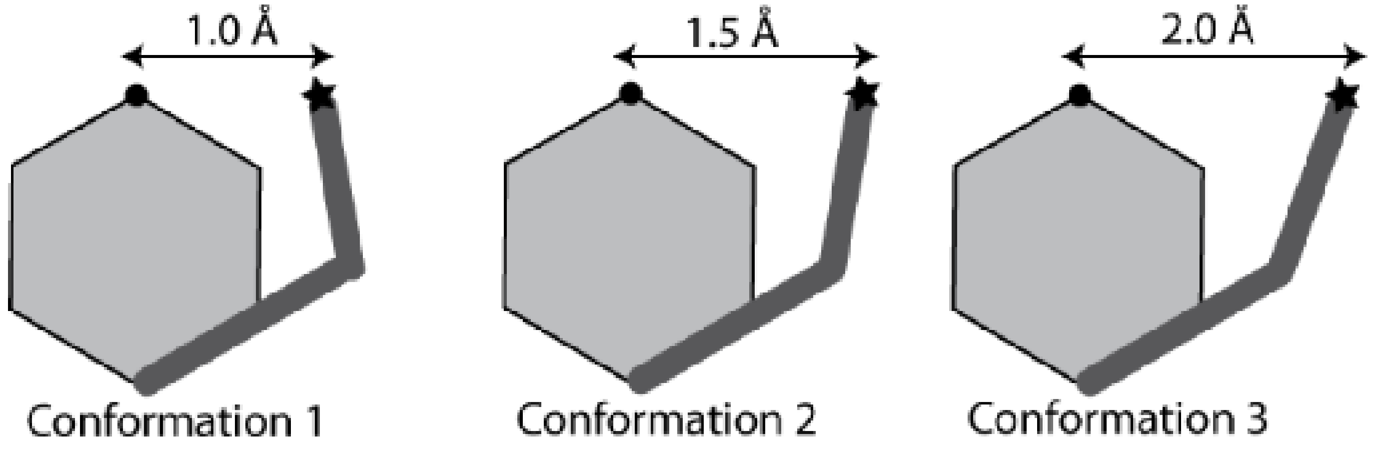
1. A peptide loop on a protein molecule was probed using fluorescence spectroscopy to measure the separation l from a point on the loop to a point on the protein as shown in figure below. After experiment you decide to model the loop as having following three different conformations:

(1) In first conformation, the loop sticks to the side of the loop with separation l=1.0 Å; and you define this as the ground state, with energy $\epsilon=0$ pN nm

(2) In second conformation, the loop is more distant from the protein with separation l=1.5 Å; you define this as first excited state with energy $\epsilon=4.14$ pN nm

(3) In third conformation, the loop is far away from the protein with separation l=2.0 Å; and you define this as second excited state with energy $\epsilon=8.28$ pN nm

Using above model, calculate following at T=300 K:



$$T = 300K \Rightarrow k_BT = \underline{4.14 pNnm}$$

(a) Partition function for the loop **(b)** Average separation i.e. $\langle l\rangle$ **(c)** Average energy of the loop $\langle\epsilon\rangle$

$$(a)\ \boxed{Z = \sum_{i=1}^{n} e^{-\epsilon_i/k_BT}},\ n=3 \Rightarrow e^{-0/k_BT} + e^{-\frac{4.14pNnm}{k_BT}} + e^{-\frac{8.28pNnm}{k_BT}}$$

$$= 1 + e^{-1} + e^{-2} \approx 1 + 0.368 + 0.135 \approx \boxed{1.503}$$

$$(b)\ \langle A\rangle = \sum_{i=1}^{n} A_i p_i \Rightarrow \boxed{\langle l\rangle = \sum_{i=1}^{3} l_i p_i} \rightarrow p_i = \frac{e^{-\epsilon_i/k_BT}}{Z} \rightarrow$$

$$l_1 = 1Å,\ l_2 = 1.5Å,\ l_3 = 2Å$$

$$\Rightarrow p_1 = \frac{1}{1.503} \approx \underline{0.665},\ p_2 = \frac{e^{-1}}{1.503} \approx \underline{0.249},\ p_3 = \frac{e^{-2}}{1.503} \approx \underline{0.090}$$

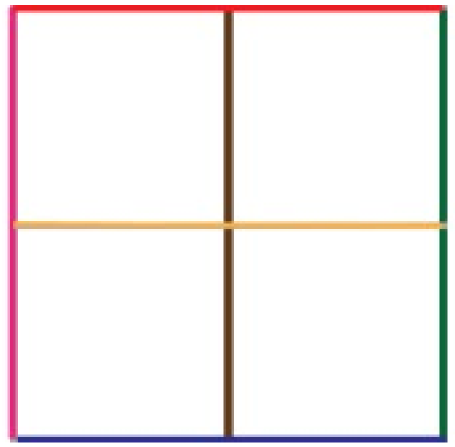
$$\langle l \rangle = 0.665 \times 1 + 0.249 \times 1.5 + 0.09 \times 2 = 0.665 + 0.374 + 0.18 \approx \boxed{1.219 \text{Å}}$$

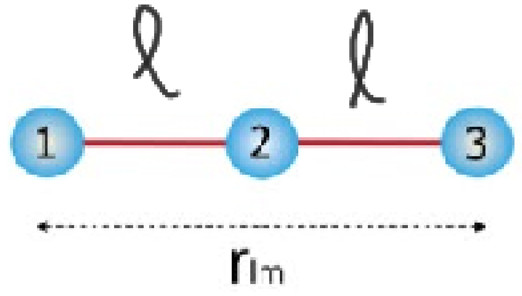
(c) $$\langle \varepsilon \rangle = \sum_{i=1}^{3} \epsilon_i p_i = 0.665 \times 0 + 0.249 \times 4.14 + 0.09 \times 8.28$$

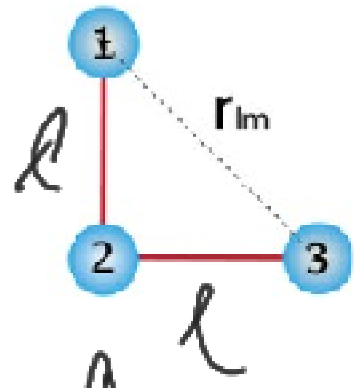
$$= 0 + 1.031 + 0.745$$

$$\epsilon_1 = 0, \epsilon_2 = 4.14, \epsilon_3 = 8.28 \qquad = \boxed{1.776 \, \text{pNnm}}$$

2. Imagine a protein made of three identical/indistinguishable connected positive charges. The length of the bond between two neighboring charges of protein 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, $U_i = \sum_{l=1}^{2} \sum_{m=l+1}^{3} \frac{A}{r_{lm}}$

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

$$= \frac{A}{r_{12}} + \frac{A}{r_{13}} + \frac{A}{r_{23}}$$

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° 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?

(a) $U_s = \frac{A}{1} + \frac{A}{2} + \frac{A}{1} = \frac{5}{2} A/nm = \boxed{2.5 k_B T}$

(b) $U_b = \frac{A}{1} + \frac{A}{\sqrt{2}} + \frac{A}{1} = \boxed{2.707 k_B T}$

↳ becouse of 90°

(c) $3+3 = \boxed{6} = W_s$

(d) $W_b = 4 \times 4 = \boxed{16}$

& (e) (f) $P_s = \frac{e^{-G_s/k_BT}}{Z}$ $\quad P_b = \frac{e^{-G_b/k_BT}}{Z}$ $\quad Z = e^{-G_s/k_BT} + e^{-G_b/k_BT}$

$$G_s = \langle U_s \rangle - TS_s = \langle U_s \rangle - k_BT \ln W_s = 2.5 k_BT - k_BT \ln 6 \approx \underline{0.708 k_BT}$$

$$G_b = \langle U_b \rangle - TS_b = \langle U_b \rangle - k_BT \ln W_b = 2.707 k_BT - k_BT \ln(16) \approx \underline{-0.066 k_BT}$$

$$Z = e^{-0.708} + e^{-(-0.066)} \approx \underline{1.56} \Rightarrow P_s \approx \frac{e^{-0.708}}{Z} \approx \boxed{0.316}, \ P_b \approx \boxed{0.684}$$

3. Show that expression for entropy $S = k_B lnW$ is equivalent to $S = -k \sum_i p_i \, lnp_i$ if k_B is replaced by k for a single macro-state assuming that all microstates have identical energies.

Start with $\boxed{S = -k \sum_i p_i \ln p_i}$

Single macrostate microstates with identical energy $\Rightarrow p_i = \frac{1}{W}$

$$= -k\left(\sum_i \frac{1}{W}\ln\frac{1}{W}\right) = -k \ln\frac{1}{W}\sum_i \frac{1}{W} = -k(-\ln W)\sum_i \frac{1}{W}$$

$$= k\ln W \times \frac{W}{W} = \underline{k\ln W}$$

Replace k back with k_B $\Rightarrow$ $\boxed{S = k_B \ln W}$

W = Number of microstate configurations in the macrostate

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

Second position $(M=4)$

$P_1(A) = 2/10 = 0.2$

$P_2(T) = 0.3$

$P_3(C) = 0.2$

$P_4(G) = 0.3$

$S_2 = -k_B\left((0.2 \ln 0.2)\times 2 + (0.3 \ln 0.3)\times 2\right)$

$= \underline{1.366 k_B}$

(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.

(i) First position $\Rightarrow n_1 = 1, N=1$ $(M=1)$ $\Rightarrow P_1 = 1$

$S_1 = -k_B p_1 \ln p_1 = \underline{0}$

Third position $(M=4)$

$P_1 = 0.1, P_2 = 0.7, P_3 = 0.1, P_4 = 0.1$

$S_3 = -k_B(0.7 \ln 0.7 + (0.1 \ln 0.1)\times 3)$

$= \underline{0.94 k_B}$

(ii) Entropy $\downarrow \Rightarrow$ Conserved $\uparrow$

First > Third > Second

most conserved

least conserved