Homework Exercise 6

1)

For this question, you will need to understand the (information-theoretic) entropy metric, its relevance to molecular sequence conservation, and its representation as sequence logos. Two good resources to begin with are an intuitive introduction to Shannon's entropy and Wikipedia (which explains, among others, how entropy is calculated for logos and how it is used for displaying the symbol distributions in each position).

A)

Given the following aligned sequences, compute the entropy at each position manually (yes, this is actually not a programming exercise). Please include the calculations in your submission (it's OK to scan hand-written answers, as long as they're legible):

ACACTT

ATGTCT

TGCCCT

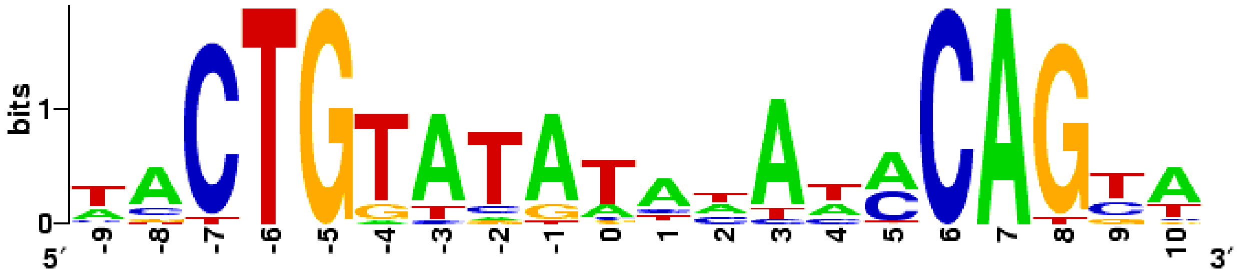
AACTAT

B)

Which positions have the highest and lowest entropies? Does that make intuitive sense?

C)

Consider the following motif logo:



Answer the following questions in one sentence:

i) Which positions are least important for the represented motif?

ii) Which letters would be the most surprising to observe at the -7 position?

2)

A)

Download the entire genome sequence of the **K-12** strain of **E. Coli** (it's an average size bacterial genome with ~5M base pairs). You can search for the RefSeq record in NCBI Genome (https://www.ncbi.nlm.nih.gov/genome).

B)

Read about the lexA repressor (in E. Coli). What is the role of this transcription factor?

C)

The logo presented in Q1C represents the binding sites of lexA repressor in the E. coli K-12 genome.

Write a regex that captures this logo (use your judgment to decide which parts are important and which aren't) and find all of its occurrences in the genome you downloaded. Don't forget to search on the negative strand as well.

How many occurrences did you find? How many would you expect to find at random?

D)

Use the sequences you found to build a 4x20 NumPy matrix with the frequency of each nucleotide at each position of the lexA repressor motif (i.e. each column in the matrix should indicate the 4 nucleotide frequencies of the sequences you found, within the relevant position, and should sum to 1).

E)

As you read in Q1, Shannon entropy is a commonly-used measure in information theory for quantifying uncertainty. Entropy of 0 indicates complete certainty, while larger values indicate growing amounts of uncertainty.

Use the function **scipy.stats.entropy** on the matrix you built to obtain a vector quantifying the amount of entropy at each position of the motif. Explain the result.

F)

Do you believe that most of the sequences you found are truly lexA repressor binding sites, or just random sequences? What future analysis could settle the issue?

3)

The following is a quiz on p-values (if you need a refresher, you can watch this video, explaining what p-values are, and this video about how to interpret them).

A) Researchers from the End Coin Bias Initiative received a new coin on their mailbox, and wanted to test if it was biased or not. They flipped the coin 20 times, counting the number of heads and tails, and ran a statistical test on the result under the null hypothesis that the coin was not biased. They got a p-value of 0.1. What does that mean? Explain in your own words.
B) What's the difference between one-tailed and two-tailed statistical tests? Which of the two options do you think is more appropriate to use in the case of coin flipping?
C) Does a p-value lower than 0.05 necessarily mean that the tested coin is biased?
D) Do you agree with the following statement: a lower p-value means stronger evidence for a biased coin.
E) The #endcoinbias campaign has gone viral on social media, putting pressure on governments to take immediate action on the issue of coin flipping fairness. To meet public demands, ECBI researchers have decided to substantially increase the number of flips in coin testing protocols. After repeatedly flipping a coin for an excruciatingly long time, they tested the result and obtained a p-value of 2.4E-19. Does it mean that the coin is likely biased?
F) If the coin is indeed biased, does such a low p-value indicate that the bias is strong?
G) Does a p-value of 0.95 indicate an unbiased coin? Is it evidence of its fairness?

Bonus Questions

1)

Python’s **urllib/urllib2** library can be used to retrieve the content of internet URLs via HTTP. Use this library to write a script that given a UniProt ID will display a short summary about the protein that includes:

- The name of the protein
- Functions
- Subcellular locations
- Keywords