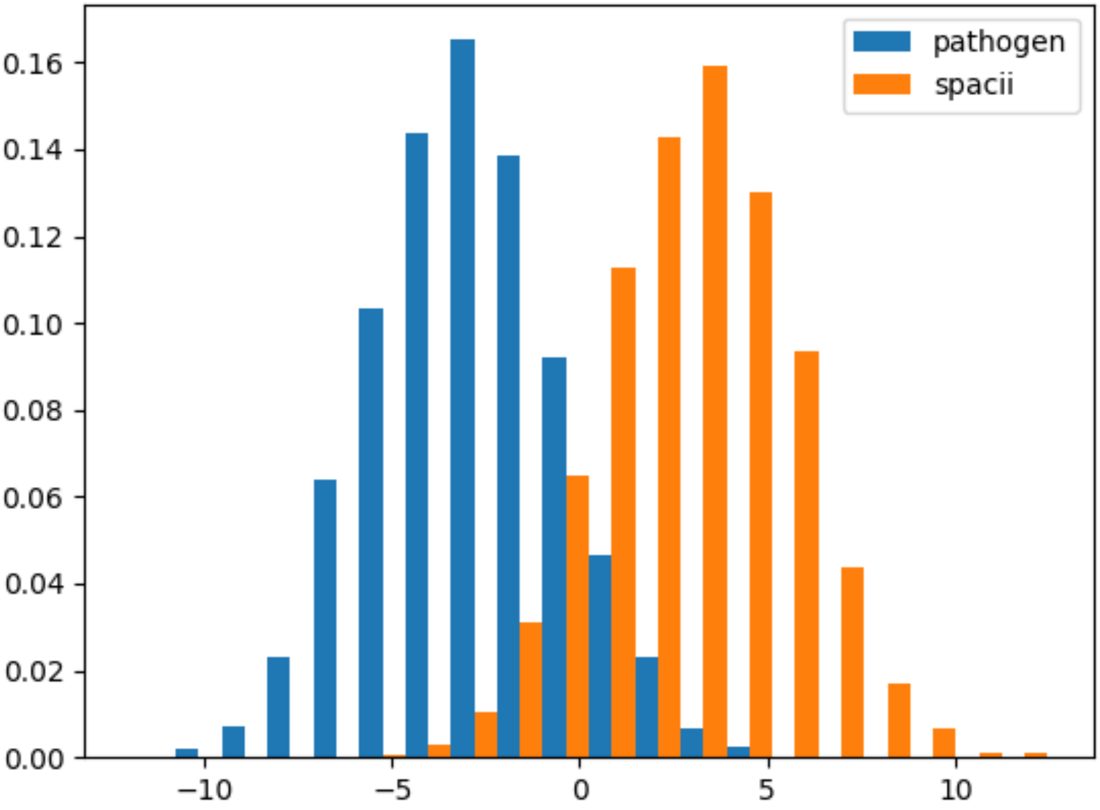
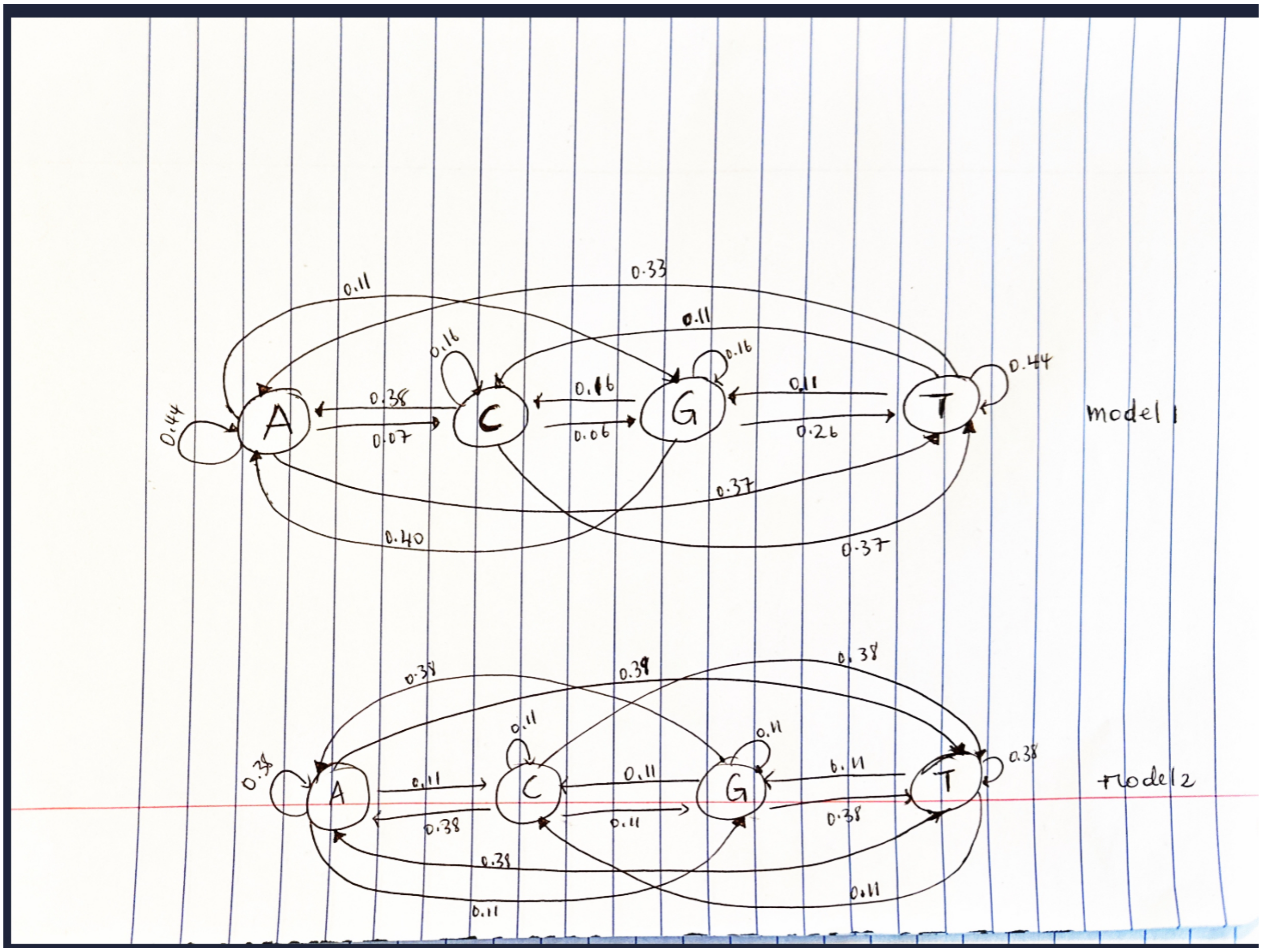
Homework 5: HMMs And Models

(100 points total)

Complete the class assignment: Nucleotide Composition HMM (50 points)



2) Draw (by hand is fine) the HMM that represents our pathogen model.



Emissions : [“A”= 0.2, “C” = 0.4 “G”= 0.2, “T” = 0.2]

3)

To combine the models into a single HMM, I'd create a new HMM with states representing both the pathogen and the Spacii nucleotides, as well as transitions based on the trained models. Dynamic programming allows us to find the most likely path through the combined HMM that maximizes the probability of the observed sequence. The merge point would be the point in the sequence where the most likely path transitions from Spacii to pathogen states, or vice versa.

Create an ORF generating HMM (50 points)

1.a)

ORF: Start Codon, Coding Region, Stop Codon.

Emissions:

- Start Codon(S)**:** ATG = 1
- Coding Region (C)**:** A = C = G= T = 0.25
- Stop Codon (T)**:** TAA=TAG=TGA = 0.33

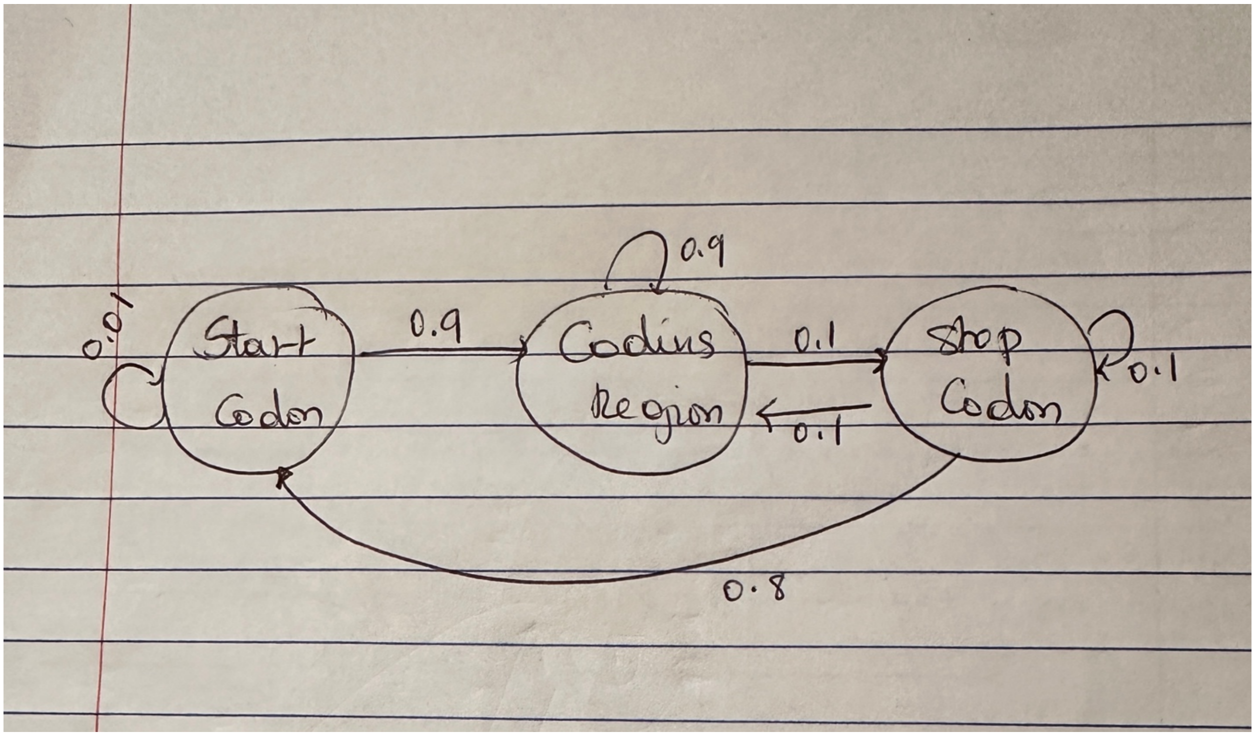
Probabilities:

A= [0.1,0.9,0.0]
[0.0,0.9,0.1]
[0.8,0.1,0.1]

k_0 = [1.0,0.0,0.0]

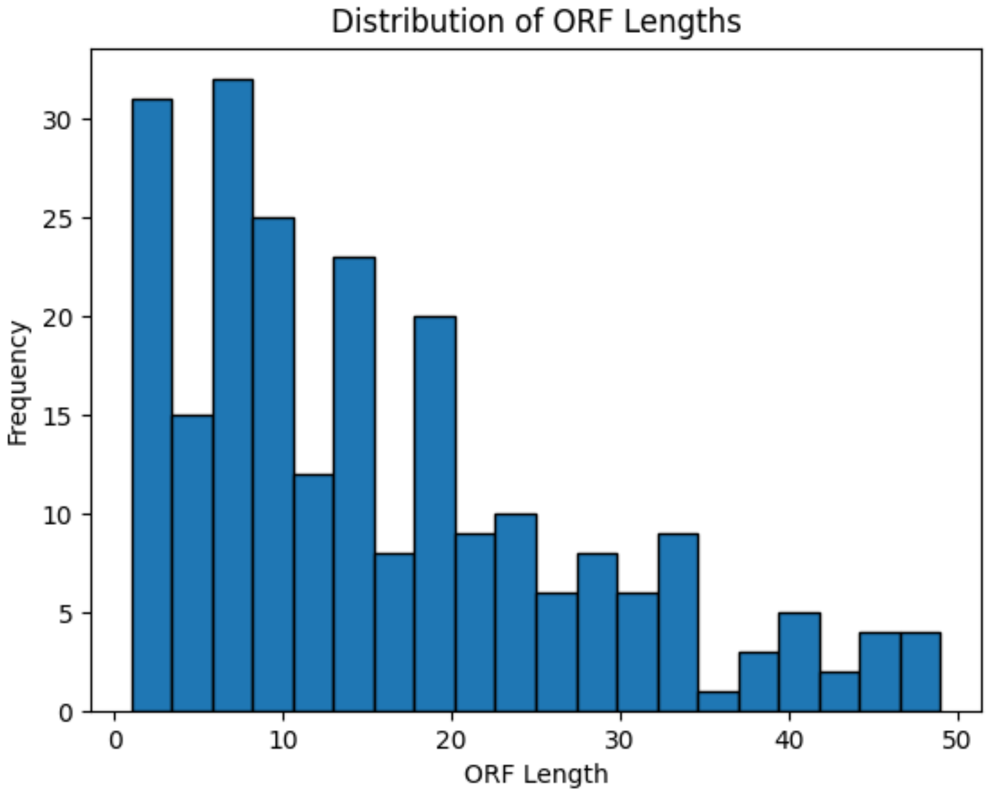
Explanation of probability values:

- Chosen to allow realistic ORF patterns in Transition probabilities(A)
- In the Coding Region, Uniform distribution in emission probabilities to ensure equal probability of emitting nucleotides A, C and G
- Start probabilities (K_0): Equal probability of starting in any state



Is your distribution what you expected based on the model parameters you've chosen? Explain. distribution is not as expected.

Because I started with a high probability most of my sequences will begin with a Start Codon. based on my transition probabilities once the model enters the Coding Region state, it is very likely to stay there This will lead to longer ORFs because there's a relatively small chance of moving to the Stop Codon from the Coding Region in any step. There's also a small chance to go from the Start Codon to the Stop Codon which would create a very short ORF. However, since this probability is quite low, the number of very short ORFs should also be low. Hence, the distribution is not as expected. my distribution should have a higher frequency of shorter ORFs and less frequent longer ORFs.


Distribution of ORF Lengths
Frequency
ORF Length