Homework 2: Motifs and Alignment
(100 points total)

Complete the class assignment (60 points)

1) Evaluate the output of your program (30 points)
    a. Do any sequences have multiple ORFs? What do you think is the most likely explanation (do you think each ORFs is a different gene)? Explain. (10 points)
        i. Yes, there are sequences that have multiple ORFs. This can occur if there are overlapping or nested ORFs. This may occur because different start codons may serve various functions and may be used for specific isoforms.
        ii. I would assume that any overlapping ORFs are from the same gene; however, ORF regions that do not overlap could represent different genes depending on the lengths of the contigs.
    b. Do any of your ORFs not end on a stop codon? Based on what you know about assembly and genomic regions that don't produce reads do you think this ORF is protein coding? (10 points)
        i. It looks like all the ORFs that I have identified all end in stop codons. Additionally, if a sequence has a low motif score, then it is less likely for the ORF to be a protein-coding and functional region.
        ii. There are a few sequences initially found with the start codon GTG that do not produce a score of over 7.5 due to its 0.5 in position 10. However, they still are probably functional regions and prompt us to reconsider whether the threshold of 7.25 may be too high in identifying ORFs.
    c. Is there a sequence that didn't have a significant hit? Look at the sequence(s) in more detail by changing some program parameters. Do you think this is a true negative or a false negative? Justify your answer. (10 points)
        i. Contigs 6 and 11 did not produce a significant hit. While there were ORF regions that were identified, they did not quite meet the 7.25 threshold. One of these sequences started with a GTG codon, achieving a high score of 5.0 while the others just did not meet the 7.25 because of the other base pair locations and the scores in the corresponding motif matrix.
        ii. These are likely to be a false negative because the program still identified it as an ORF. While it was unsatisfactory for the given parameters, that is not necessarily a sign of low functionality for the sequence.

Consider the following about finding non-real ORFs (20 points)

2) Let's assume that we have a random sequence in which we are trying to identify an ORF in a genome that's 50% GC.

a. Once we have identified a start, what is the probability that the next codon is a stop codon? (5 points) Hint: in a random sequence this is true about the probabilities.

$$P_{AAA} = P_{GCA} = P_{TTT}$$

i. In a random sequence, if there is a 50% chance of the nucleotide being a G or C, then there is alternatively a 50% chance it could be an A or T. There are three potential stop codons (TAA, TGA, and TAG). The probability of each of these are calculated as follows:
   1. P(TAA) = P(T) x P(A) x P(A) = ¼ x ¼ x ¼ = 1/64
   2. P(TGA) = P(T) x P(G) x P(A) = ¼ x ¼ x ¼ = 1/64
   3. P(TAG) = P(T) x P(A) x P(G) = ¼ x ¼ x ¼ = 1/64
   4. P(Stop Codon) = P(TAA) + P(TGA) + P(TAG) = 3/64 x 100 = 4.69%

b. What is the probability that we have 600 bases (200 codons) beyond the start codon with NO stop codon encountered? (5 points). Hint: You are sampling 200 times.

i. From our previous problem, we determined that there is a 3/64 chance of identifying a stop codon. So, on the other end, there is a 61/64 chance of not encountering one. If we decide we want to sample 200 times. Then we must raise this probability to the 200th power.

ii. Therefore, the probability of not identifying a stop codon in 600 bases comes out to (61/64) ^200 = 0.0067%, which is very slim.

c. Calculate how many stops we expect to see in 200 codons. (5 points)

i. To calculate this, we take the probability of seeing a stop codon, which was 3/64, and multiply that by 200. As a result, we get 9.375, which means that we most likely will come across 9 stop codons if we have a sequence that is 200 codons long.

d. If the number of stops we observe is Poisson distributed (it is): What is the probability that we see 0 stop codons? Use the Poisson distribution. Show your work. (5 points). Hint: You calculated lambda in part c.

i. Using the Poisson distribution formula, we want to plug in the values accordingly. In this case, the lambda value would be 9.375. In the denominator, we would replace *k* with 0. This equation simplifies to e^-9.375 which results in a percentage of 0.00848%. This is still extremely small, just like the value we got in part b.

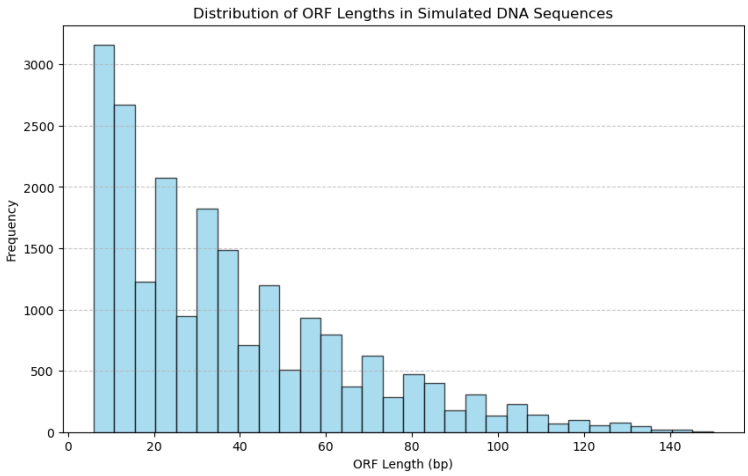
e. How does this compare to your answer in part b? (0 points but cool!) Part b represents the calculation from the binomial distribution. It's pretty satisfying

that different ways of looking at the problem yield similar answers! Part b requires knowing the exact probability though, while d just relies on an estimate of lambda so can be more broadly applicable.

i. The values are very close to each other, both significantly small. This indicates that the Poisson distribution is a good approximation.

Consider simulating sequences (20 points)

In Bioinformatics we often want to check our thinking with simulating the data. I have provided some code for generating a random sequence of bases (generateRandomSequences.py) in the Canvas Module.



3) Use this code to simulate some sequences. Treat any ATG as start and any stop codon as a stop. Ignore any ORF that reaches the end of the sequence before a stop codon.
   a. What is the distribution of ORF lengths across your sequences? Make sure you have enough sequences so that you think your distribution is reasonable. Include a visualization of your choice. (10 points)
      i. The histogram shows the distribution of ORF lengths across 15,000 randomly generated sequences of length 150 nucleotides. The distribution is extremely positively skewed with a right tail.
      ii. There are a few reasons this may be the case. For a sequence to have a longer ORF, then the start codon ATG must begin earlier in the sequence. In randomly generated artificial sequences, there is no biological placement of ATG so there are less chances of finding a long ORF downstream.
      iii. Additionally, the smaller the sequence length, the greater the chance that the next codon is a stop. Longer ORFs seem to be rarer.
   b. What is the fraction of randomly generated contigs of size 150 nucleotides that contain ORFs > 60 nucleotides. You may approximate from part (a) or generate

the data by creating sequences of size 150 and counting how many have long ORFs. Explain how you arrived at your answer (5 points)

i. The fraction of sequences that contain at least one ORF longer than 60 nucleotides is approximately 22.1%. This was determined by the latter approach of counting the number of sequences in which an ORF of at least 60 nucleotides appeared and dividing by the total number of sequences generated (15,000).

c. Let's say you wanted to change your code from the class assignment (you don't actually need to do this) to return all ORFs with an ATG start and a stop codon end. Describe how you could do this by changing only global variables. It's ok if your resulting program misses some edge cases like ATGs very early in the sequence or returns extra hits like ORFs terminating at sequence end. (5 points)

i. The changes could modify the scanSeq() function to ensure that only ORFs starting with ATG and ending with a stop codon (TAA, TAG, or TGA) are identified. Instead of explicitly checking for multiple start codons (ATG or GTG), we could introduce a new global variable that is set to ATG. This limits ORF identification to only those beginning with ATG.

ii. Similarly, a global variable for stop codons could be defined to store valid stop codons. The function then references these variables during scanning and motif scoring.

iii. This approach may not catch nested ORFs or sequences that do not end with a stop codon.