Homework 2: Motifs and Alignment

(100 points total)

Assignment guidelines

- Submit your assignment files on canvas under module 4: Motifs and Computer Set-up
- Please submit your code in a file called [name].py. Your code should be easy to open in a text editor so that someone can download and use the function you write.
- Please submit a pdf with the answers to the questions at the bottom of the assignment (and your visualization)
- Please submit a text file with the output of your code called [name].fa

Complete the class assignment (60 points)

1) Create a function called "identify ORFs" which takes as input a file with contigs outputs a file with a FASTA sequence of contigs with labels which include what contig they came from (ex: sequence 1_ORF1) (30 points)
    a. Code meets specifications – the function exists and makes correct input and output (10 points)
        i. Code includes scanSeq() function which takes in a sequence and returns an array of sequences, an array of corresponding ORF start positions, and an array of corresponding ORF lengths
        ii. Code includes scoreMotif() functions which takes in a 13bp sequence and returns the score
        iii. Use parameters
            1. x >7.25 as quality score cutoff
            2. 60 Nucleotides coding for amino acids as min ORF length
            3. Motif matrix found in MotifCode.txt
        iv. Format output as:

> contig x|length of putative ORF2|at pos y /
> sequence of putative orf2
> > contig 2|Length 62|at position 15
> ATG….TTC…TAA
> > contig 4|Length 62|at position 12
> ATG….GGG…TAA
> > contig 4|Length 62|at position 54
> ATG….GGG…TAA
> > contig 5|Length 77|at position 23
> GTG….ATC…TGA

    b. All returned sequences start with ATG or GTG (10 points)
    c. All ORFs over threshold are found (10 points)

2) 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)

```
secretSeq_0|test: One ORF detected
secretSeq_1|test: One ORF detected
secretSeq_2|test: One ORF detected
secretSeq_3|test: Multiple ORFs detected (3 ORFs)
secretSeq_4|test: Multiple ORFs detected (2 ORFs)
secretSeq_5|test: One ORF detected
secretSeq_6|test: One ORF detected
secretSeq_7|test: One ORF detected
secretSeq_8|test: One ORF detected
secretSeq_9|test: One ORF detected
secretSeq_10|test: No ORFs detected
secretSeq_11|test: One ORF detected
secretSeq_12|test: One ORF detected
secretSeq_13|test: One ORF detected
secretSeq_14|test: Multiple ORFs detected (3 ORFs)
secretSeq_15|test: One ORF detected
secretSeq_16|test: One ORF detected
secretSeq_17|test: Multiple ORFs detected (2 ORFs)
secretSeq_18|test: One ORF detected
secretSeq_19|test: One ORF detected
```

An open reading frame (ORF) is a segment of RNA that can be continuously translated by ribosomes into proteins. It consists of codons—groups of three nucleotides—each coding for a specific amino acid. An ORF extends until a stop codon is encountered, which signifies the end of protein synthesis. Longer ORFs are more likely to represent functional genes. Additionally, due to the two strands of DNA, each strand can produce three reading frames, resulting in a total of six potential ORFs for each DNA segment (Shchelochkov, 2019).

The results show that many sequences have several open reading frames (ORFs), indicating different biological processes like alternative translation initiation, overlapping genes, or varied functions among the ORFs found. For example, secretSeq_3 contains three ORFs, while secretSeq_4 has two and so on. This could happen if one mRNA transcript creates different protein isoforms or gene duplication leading to similar sequences with distinct functions. However, having multiple ORFs does not mean each one is a separate gene. Further experimental validation, such as expression analysis or functional assays is necessary to understand pattern of multiple ORFs.

Factors that lead to multiple ORFs include alternative initiation, where a single gene produces multiple protein isoforms due to different start codons; overlapping genes, where different ORFs share nucleotide sequences; readthrough transcription, which may lead to additional ORFs beyond a stop codon; and gene duplication, resulting in distinct but similar ORFs. To clarify the relationships between the ORFs and their potential biological roles, further investigation, such as comparative genomics, is required.

3) 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)

The detection of potential protein-coding sequences is largely dependent on the presence of a stop codon at the end of an open reading frame (ORF). Stop codons act as signals for the termination of translation, indicating that the ORF is likely to produce a functional protein. In our analysis, we confirmed that each ORF ends with a stop codon, which supports the hypothesis that these sequences could be protein-coding.

While all identified ORFs terminate with a stop codon, the absence of read data in certain genomic regions raises important considerations. Regions lacking reads may indicate low expression levels, technical biases, or structural variations that that complicate sequencing. In conclusion, while all our ORFs terminate with stop codons, the actual protein-coding potential of these ORFs relies on other factors, such as expression levels and experimental validation. An ORF in a read-deficient area might not be actively translated, suggesting that further investigation is necessary to determine its functional status.

4) 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)

At first, when we set the parameters with a minimum ORF length of 60 and a cutoff score of 7.25, the sequences `secretSeq_5|test` and `secretSeq_10|test` didn't show any significant hits. However, after adjusting the parameters to a minimum ORF length of 30 and a cutoff score of 5.0, all sequences showed significant hits. This observation suggests that the initial settings were too stringent, potentially leading to false negatives. By broadening the criteria, we were able to find ORFs that we didn't catch before, hinting that these sequences might have coding potential that we initially overlooked. Therefore, it is reasonable to conclude that the lack of significant hits in the initial analysis represents a false negative, rather than a true negative, highlighting how crucial it is to optimize parameters in bioinformatics to accurately spot potential protein-coding sequences.

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

3) 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)

i. Hint: in a random sequence this is true about the probabilities

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

The Likelihood of the Next Codon Being a Stop Codon :

In a random sequence, the three stop codons (TAA, TAG, TGA) have an equal likelihood of occurrence. To ascertain the probability of any given codon being a stop codon, we proceed as follows:

Total number of codons: There are 64 possible codons, derived from the combination of 4 bases multiplied by themselves three times ($4 \times 4 \times 4 = 64$).

Number of stop codons: There are 3 designated stop codons (TAA, TAG, TGA).

P(stop codon) = Number of stop codons/ Total number of codons = 3/ 64

So, the probability that the next codon following the start codon is a stop codon is approximately 0.046875 or (4.69%).

3.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.

Probability of Not Encountering a Stop Codon Over 600 Bases

The probability of not encountering any stop codons within 200 codons (which corresponds to 600 bases), following a binomial distribution.

The probability of not encountering a stop codon in a single codon is determined as follows:

P(not a stop codon) = 1 - P(stop codon) = 1 - (3/64) = (61/64)

Our goal is to find the probability of this event occurring across 200 codons:

P(no stop codon in 200 codons) = (P(not a stop codon))^200 = (61/64)^200

We can now calculate this value:

P(no stop codon in 200 codons) ≈ (0.953125)^200 ≈ 0.01487

Therefore, the probability of not encountering any stop codons over 600 bases (or 200 codons) is approximately 0.01487, which is about 1.487%.

3. c

Calculate how many stops we expect to see in 200 codons. (5 points)
From prior calculations, the probability of a codon being a stop codon is: 4.69%

The expected number of stop codons (E) in each number of trials can be calculated using the formula:

$$\mathrm{E} = \mathrm{n} \times \mathrm{P}(\text{stop codon})$$

Where, n = total number of codon (200 in this case), P = probability of stop codon

$$\mathrm{E} = 200 \times \frac{3}{64} \approx 200 \times 4.69 \approx 9.38$$

It is anticipated that there will be approximately 9.38 stop codons present within a sequence of 200 codons. As it is common practice to round values to whole numbers in the context of counting, it is reasonable to estimate that there will be around 9 or 10 stop codons.

---

3. 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.

To find the probability of observing 0 stop codons using the Poisson distribution,

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where,

$P(X = k)$ is the probability of observing k events (stop codons) in a fixed interval,
λ is the expected number of events (stop codons),
k is the actual number of events observed (in this case, k = 0),
e is the base of the natural logarithm (approximately equal to 2.71828).

Calculation:

$\lambda \approx 9.375$ for 200 codon, k = 0 in the poisson formula:

$$P(X = 0) = \frac{9.375^0 e^{-9.375}}{0!} = \frac{1 \times 0.0000836}{1} \approx 0.0000836$$

The probability of 0 stop codons in 200 codons, using the Poisson distribution, is approximately 0.0000836, or 0.00836%.

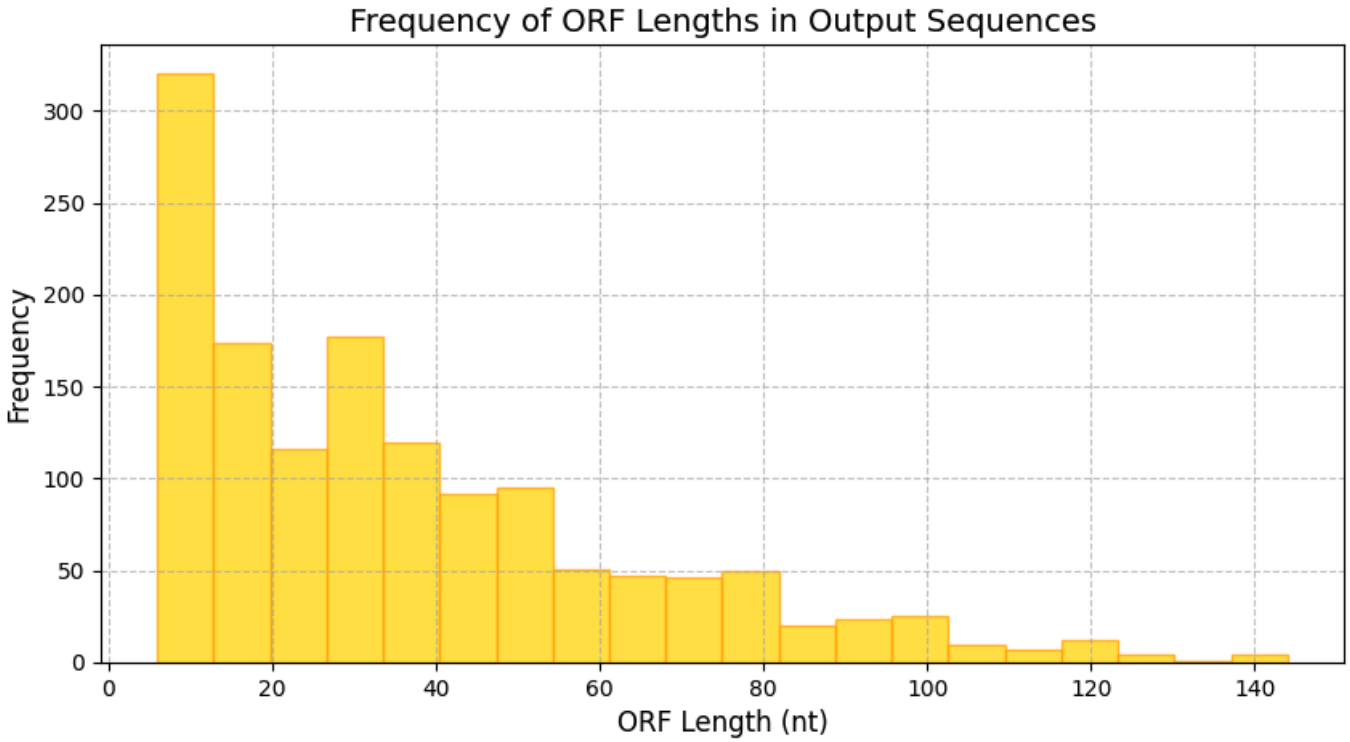
3.e How does this compare to your answer in part b? (0 points but cool!)

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.

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.

5) 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 that you think your distribution is reasonable. Include a visualization of your choice. (10 points)



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)

   Fraction of sequences with ORFs longer than 60 nucleotides: 0.23

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)

   To effectively identify open reading frames (ORFs) that start with the codon "ATG" and end with a stop codon, it's recommended to use global variables and lower the cut-off score to include all sequences initiated by "ATG." Additionally, adjusting the minimum ORF length parameter to 1 will allow for the detection of shorter ORFs, broadening the range of identified sequences without compromising the specificity of the start codon.

   This approach enables a comprehensive exploration of potential ORFs, providing valuable insights into the genetic landscape under study, especially in exploratory research or less characterized genomic regions.