CSC 314**, Bioinformatics Lab #12: Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Gene Prediction**

This lab includes concepts related to Gene Prediction, namely, sequence translation and probability concepts such as Bayes’ Theorem.

Sequence Translation Questions

Consider the sequence below:

5’ – ACGGATC – 3’

1. If the above sequence is on the *sense* strand, complete the table below to show the three reading frames. Follow the format of page 6 in the course notes, <https://gdancik.github.io/CSC-314/data/notes/GenePrediction.pdf>.

|  |  |  |  |
| --- | --- | --- | --- |
| Reading frame 1 |  |  |  |
| Reading frame 2 |  |  |  |
| Reading frame 3 |  |  |  |

1. If the above sequence is on the *template* strand, complete the table below to show the three reading frames.

|  |  |  |  |
| --- | --- | --- | --- |
| Reading frame 4 |  |  |  |
| Reading frame 5 |  |  |  |
| Reading frame 6 |  |  |  |

1. Use the *Expasy Translate* tool (<https://web.expasy.org/translate/>) to translate the above sequence, setting the Output format to “Includes nucleotide sequence”. Take a screenshot of your results showing all 6 reading frames and the corresponding amino acids, and paste the screenshot here. Note that the 6 reading frames should correspond to your results from questions (1) and (2).
2. The file *sequences.fasta* contains 3 DNA sequences in FASTA format. Use the *Expasy Translate Tool* to translate the sequences and complete the tables below.

Note: an open reading frame (ORF) in DNA begins with an ATG (the start codon) and extends to the first stop codon (but does not include the stop codon). If no stop codon is found, the ORF extends to the end of the sequence. An ORF specified as a polypeptide will therefore begin with a Methionine (an M) and extend to the first stop codon, or to the end of the sequence if no stop codon is present. A sequence such as **M**E**M**A contains *two* open reading frames: MEMA and MA.

Note: the *frame* is given by Expasy and will be 5’3’ Frame 1, 5’3’ Frame 2, etc.

**Open Reading Frames in Sequence 1**

|  |  |  |  |
| --- | --- | --- | --- |
| Frame | Translation | Length | Followed by a StopCodon (Yes/No) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Open Reading Frames in Sequence 2**

|  |  |  |  |
| --- | --- | --- | --- |
| Frame | Translation | Length | Followed by a StopCodon (Yes/No) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Open Reading Frames in Sequence 3**

|  |  |  |  |
| --- | --- | --- | --- |
| Frame | Translation | Length | Followed by a StopCodon (Yes/No) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. Based on your answers above (4), complete the table below to specify the *sequence number*, *frame, translation*, and *length* for predicted genes that meet the following requirements.
	1. A start codon and stop codon must exist in the same reading frame
	2. The ORF must be at least 5 amino acids

**Predicted Genes**

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence Number | Frame | Translation | Length |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Probability Questions

Use the table below (the same one from your notes) to answer questions 1 - 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class status** | **M** | **F** | **Total** |
| Soph | 4 | 2 | **6** |
| Junior | 3 | 2 | **5** |
| Senior | 3 | 1 | **4** |
| **Total** | **10** | **5** | **15** |

1. If a student is selected at random, what is the probability that they are a senior?
2. If a student is selected at random, what is the probability that they are a female?
3. The probability that a randomly selected student is a *female* AND a *senior* is the number of female seniors (which is 1) divided by the total number of students (which is 15). Calculate this probability using the formula below:

$P\left(female and sophomore\right)= P\left(senior\right)×P(senior)$.

1. Show that $P(male|junior)∝$ 2 / 15

Note: you need to use the formula $P\left(B\right) ∝P(B|A)×P(A)$.

1. Approximately 12% of females will develop invasive breast cancer in their lifetime (and 88% of females will not). For females that have invasive breast cancer, a mammogram will detect the cancer (be positive) about 40% of the time. However, a female that does not have breast cancer will have a positive mammogram about 6% of the time.

We know that $P\left(B\right) ∝P(B|A)×P(A)$.

Let *Br* = Breast Cancer, *N* = No breast cancer, and use *+* for a positive test.

Complete the following:

* 1. $P\left(Br\right)=$

* 1. $P\left(Br\right)= $
	2. $P\left(N\right)= $
	3. If a female tests positive, show that it is actually more likely that they do *not* have breast cancer than do. Specifically, calculate $P\left(+\right)$ and $P\left(N\right)$ using the formula above.

Note: the calculation above, which takes into account the likelihood of a *false positive*, is one aspect used to determine appropriate recommendations for mammogram screenings. This question is based on recommendation changes made in 2015 (<https://fivethirtyeight.com/features/science-wont-settle-the-mammogram-debate/>), which caused some controversy. These recommendations have more recently been updated, partly over slight increases in cancer rates and disparities in cancer diagnoses in minorities – see <https://www.bcrf.org/uspstf-new-breast-cancer-screening-guidelines-2023/>.