CSC 314, Lab #13, Spring 2025 Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

HMM and Galaxy

**Part I**

The HMM below models gene regions (G) and non-gene regions (N), based on the fact that genes have higher GC content (guanine and cytosine nucleotides) than non-gene regions. All probabilities are on the log2 scale.



1. What is the initial state probability for “G”, $P(G)$? Give the probability, and not the probability on the log2 scale (rounded to 3 decimal places)
2. What is the transition probability $P(G\_{i+1}|N\_{i})$? Give the probability, and not the probability on the log2 scale (rounded to 3 decimal places).

Suppose the following sequence is observed: **aaggc**. Use the matrix below to answer the remaining questions in this section.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **a** | **a** | **g** | **g** | **c** |
| Gene (G) | -4.4 |  -7.21 | -8.03 |  |  |
| Non-Gene (N) | -2.21 | -4.23 | -6.85 |  |  |

1. What is the optimal (predicted) gene structure for the dinucleotide sequence *aa*? The probability of that structure (given *aa*) is proportional to what value?
2. Complete the above dynamic programming matrix (you must show your work)
3. What is the optimal (predicted) gene structure for the nucleotide sequence *aaggc*?
4. Suppose that the same HMM is used to analyze a sequence that is 1000 nucleotides long. The number of possible hidden states is an extremely large number that begins with a 1. How many digits are in this number?
5. Using the Viterbi algorithm, how many probability calculations are made when finding the optimal hidden state sequence (Note: you can assume that each state requires the same number of calculations).

**Part II**

1. Use Galaxy (<https://usegalaxy.org>) to find the number of genes on each chromosome, not including pseudogenes. Paste the links below for the following:
	1. your history analyzing the partial gene list (human\_genes\_partial.tsv)
	2. your workflow for counting the number of genes on each chromosome
	3. your history analyzing the full gene list (human\_genes.tsv)