Diego Segovia

HMM Lab

Part I. Mammograms and Breast Cancer

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 (will 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.

N = Normal

B = Breast Cancer

+ = positive mammogram

$P(N) = 0.88$ $\quad$ $P(+|B) = 0.40$

$P(B) = 0.12$ $\quad$ $P(+|N) = 0.06$

1. Find P(+) = probability a randomly selected female would have a positive mammogram.
   10.08%

   $P(+) = P(+|B) \cdot P(B) + P(+|N) \cdot P(N)$

   $P(+) = (0.40)(0.12) + (0.06)(0.88)$

   $P(+) = 0.1008$

2. Calculate P(Br|+) = the probability that a female with a positive mammogram has breast cancer. Based on this result, is it likely that an individual with a positive mammogram has breast cancer?
   47.62%
   Based on this result, it is not very likely that an individual with a positive mammogram has breast cancer, as the probability is less than 50%.

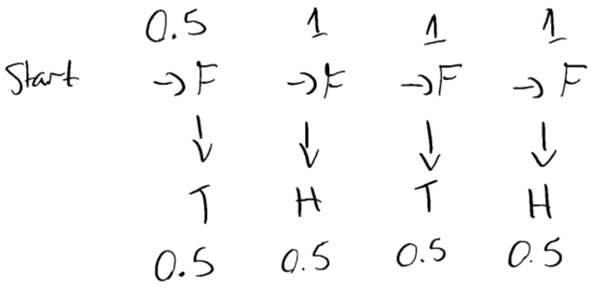
$$P(B|+) = \frac{P(+|B)P(B)}{P(+)} = \frac{(0.4)(0.12)}{0.1008}$$

$$= 0.4762$$

Part II.

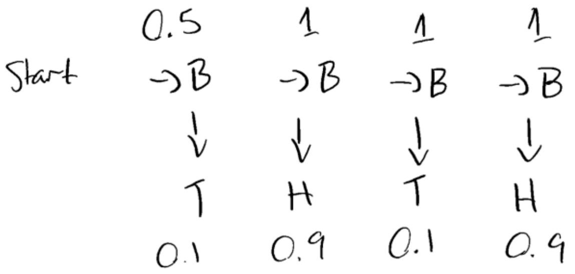
Consider the HMM on the right which models the selection of a single coin that is then tossed multiple times. Suppose the following sequence is observed from selecting and flipping a coin 4 times: **THTH**

1. Given this observation of THTH, the probability that the fair coin was selected is proportional to what value? **0.03125 (3.125%)**



$$(0.5)^5 (1)^3 = 0.03125$$

2. Given this observation of THTH, the probability that the biased coin was selected is proportional to what value? **0.00405 (0.404%)**



$$(0.5)(0.1)^2 (0.9)^2 (1)^3 = 0.00405$$

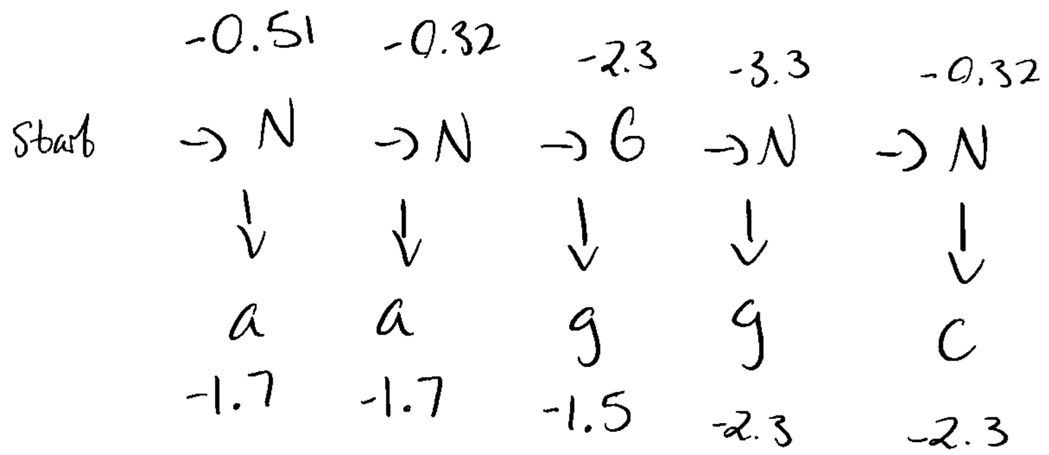
3. Given this observation of THTH, how many times more likely is it that the fair coin was selected than the biased one? **7.72 times**

$$\frac{0.03125}{0.00405} \approx 7.72 \text{ times}$$

Part III.

Consider the HMM on the right, which models gene regions (G) and non-gene regions (N) in the genome, based on the fact that genes have higher GC content (guanine and cytosine nucleotides) than non-gene regions. Suppose the following sequence is observed: **aaggc**

Given this observation of aaggc, you will show that the probability of the hidden state sequence NNGNN is proportional to $\mathbf{2^{-16.25}}$.



$$2(-1.7)+3(-2.3)+2(-0.32)-1.5-3.3-0.51$$
$$= -16.25$$

Part IV.

	a	a	g	g	c
Gene (G)	-4.4	-7.21	-8.03 →	-9.68 →	-11.33
Non-Gene (N)	-2.21 ↗ →	-4.23 ↗ →	-6.85 →	-9.47 →	-12.09

a → a → g → g

G → G : -8.03 + (-0.15 -1.5) = -9.68 * G
N → G : -6.85 + (-2.3 -1.5) = -10.65

G → N : -8.03 + (-3.3 -2.3) = -13.63
N → N : -6.85 + (-0.32 -2.3) = -9.47 * N

a → a → g → g → c

G → G : -9.68 + (-0.15 -1.5) = -11.33 * G
N → G : -9.47 + (-2.3 -1.5) = -13.27

G → N : -9.68 + (-3.3 -2.3) = -15.28
N → N : -9.47 + (-0.32 -2.3) = -12.09 * N

1. What is the optimal gene structure for the dinucleotide sequence aa? **NN**
   The probability of that structure (given aa) is proportional to what value? **-4.23 (5.33%)**
2. Complete the above dynamic programming matrix.
3. What is the optimal gene structure for the nucleotide sequence aaggc? **NNGGG**
4. The probability that the optimal gene structure produced the sequence aaggc is proportional to what value? **-11.33 (0.039%)**

Part V.

1. Suppose that this same HMM was used to analyze a sequence 30 nucleotides long. From the choices below, approximately how many possible hidden state sequences are there.
   a. 1 thousand
   b. 1 million
   c. 1 billion
   d. 1 trillion
2. Using the Viterbi algorithm, how many probability calculations are made when finding the optimal hidden state sequence? $\mathbf{4^2 * (30 - 1) = 464}$
3. How does the Viterbi algorithm compare to a "brute force" approach that would require finding the probability of every possible state sequence? **The brute force would require finding 4^{30} possibilities compared to the Viterbi algorithm's 464 possibilities making the Viterbi algorithm a substantially faster option.**