Part I. Mammograms and Breast Cancer

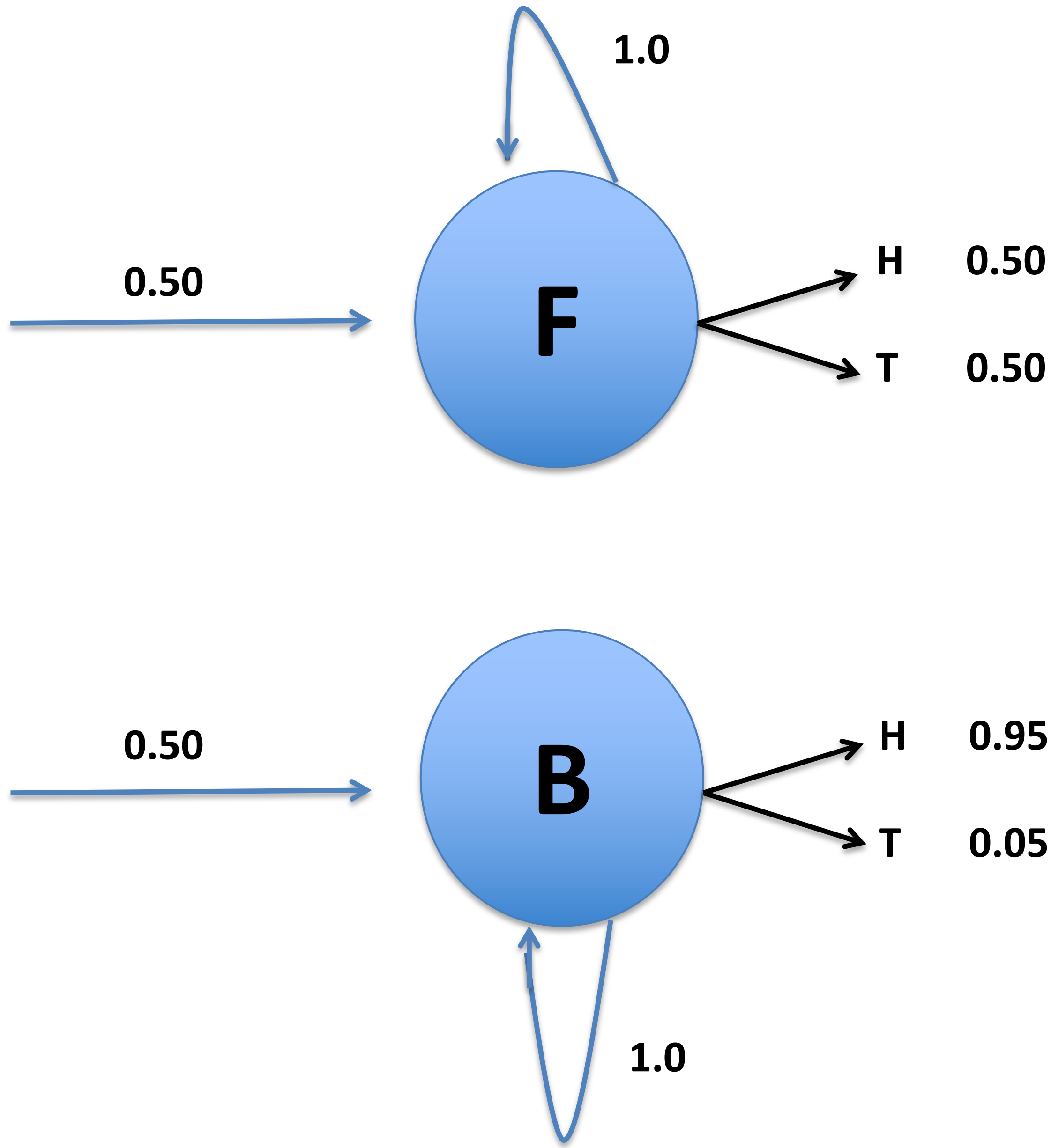
- Approximately 12% of females will develop invasive breast cancer in their lifetime (and 88% of females will not):
- For females that have invasive breast cancer, a mammogram will detect the cancer (will be positive) about 40% of the time:
- However, a female that does not have breast cancer will have a positive mammogram about 6% of the time:
- We know that $P(A|B) \propto P(B|A)*P(A)$
- Let Br = Breast cancer, N = No breast cancer, and use + for a positive test

1. What is $P(Br)$, $P(+|Br)$, and $P(+|N)$?
2. If someone has a positive mammogram, which is more likely, that they have breast cancer, or that they don't? Specifically, find $P(Br|+)$ and $P(N|+)$ using the formula above, to answer the question: if someone tests positive, how much more likely are they to NOT have breast cancer than have it?

- Note: this question provides insight into why the U.S. Preventive Services Task Force (USPSTF) advised against routine mammogram screening until women are 50 years old. More information: http://fivethirtyeight.com/features/science-wont-settle-the-mammogram-debate/. Follow-up: the USPSTF has changed its recommendations, partly over slight increased cancer rates and disparities in cancer diagnoses in minorities – see https://www.bcrf.org/uspstf-new-breast-cancer-screening-guidelines-2023/.

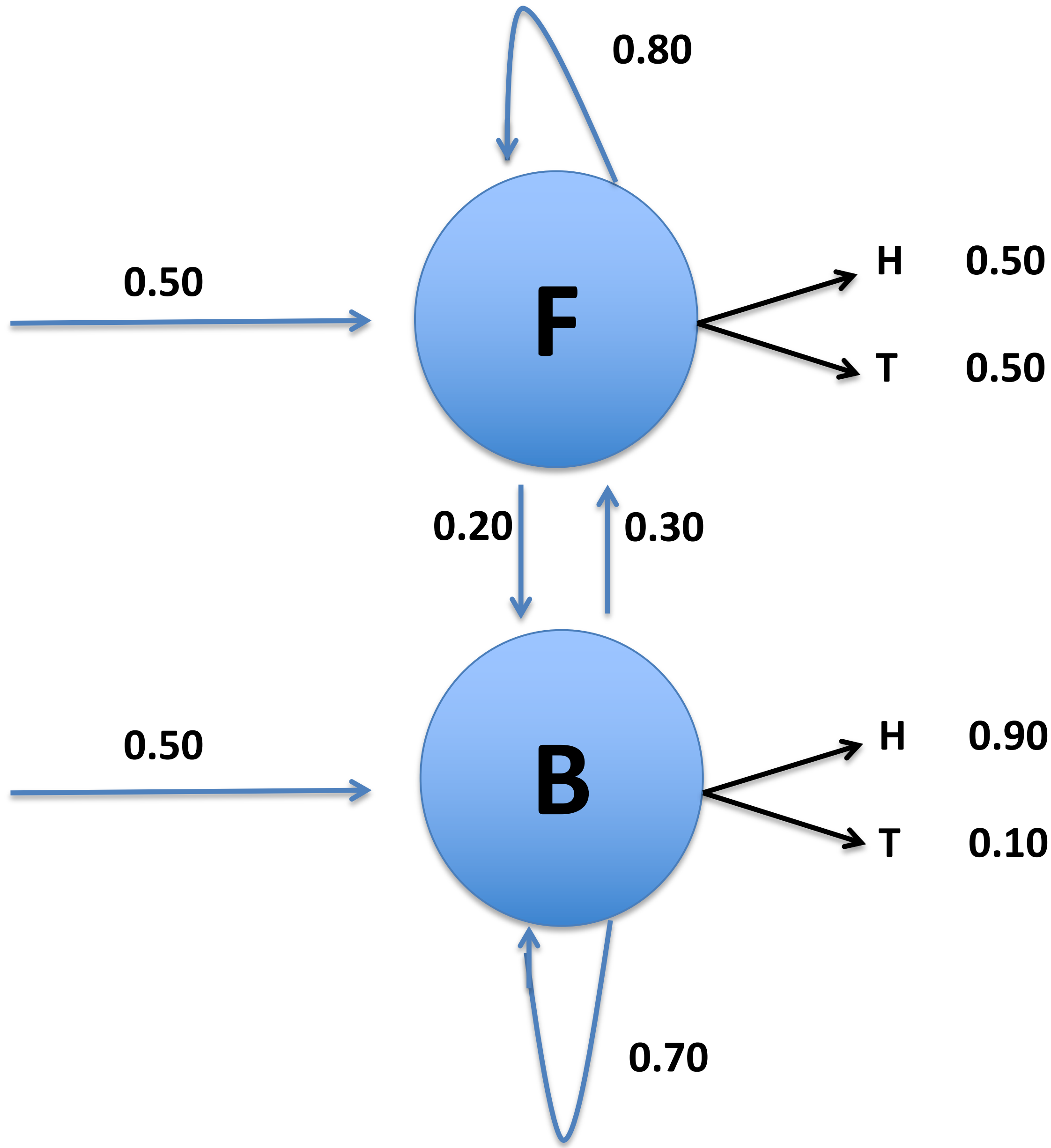
Part II. Consider the HMM on the right which models the selection of a single coin that is then tossed multiple times. Suppose the following sequence is observed from selecting and flipping a coin 4 times: **THTH**

1. Given this observation of *THTH*, the probability that the fair coin was selected is proportional to what value?
2. Given this observation of *THTH*, the probability that the biased coin was selected is proportional to what value?
3. Given this observation of *THTH*, how many times more likely is it that the fair coin was selected than the biased one?



Part III. Consider the HMM on the right which models the selection of a one or more coins that are tossed. Suppose the following sequence is observed: **THTH**

1. Given this observation of *THTH*, the probability that sequence of coins was FFFF is proportional to what value?
2. Given this observation of *THTH*, the probability that the sequence of coins is FBFB is proportional to what value?
3. Given this observation of *THTH*, is it more likely that the sequence of coins was FFFF or FBFB? How much more likely?

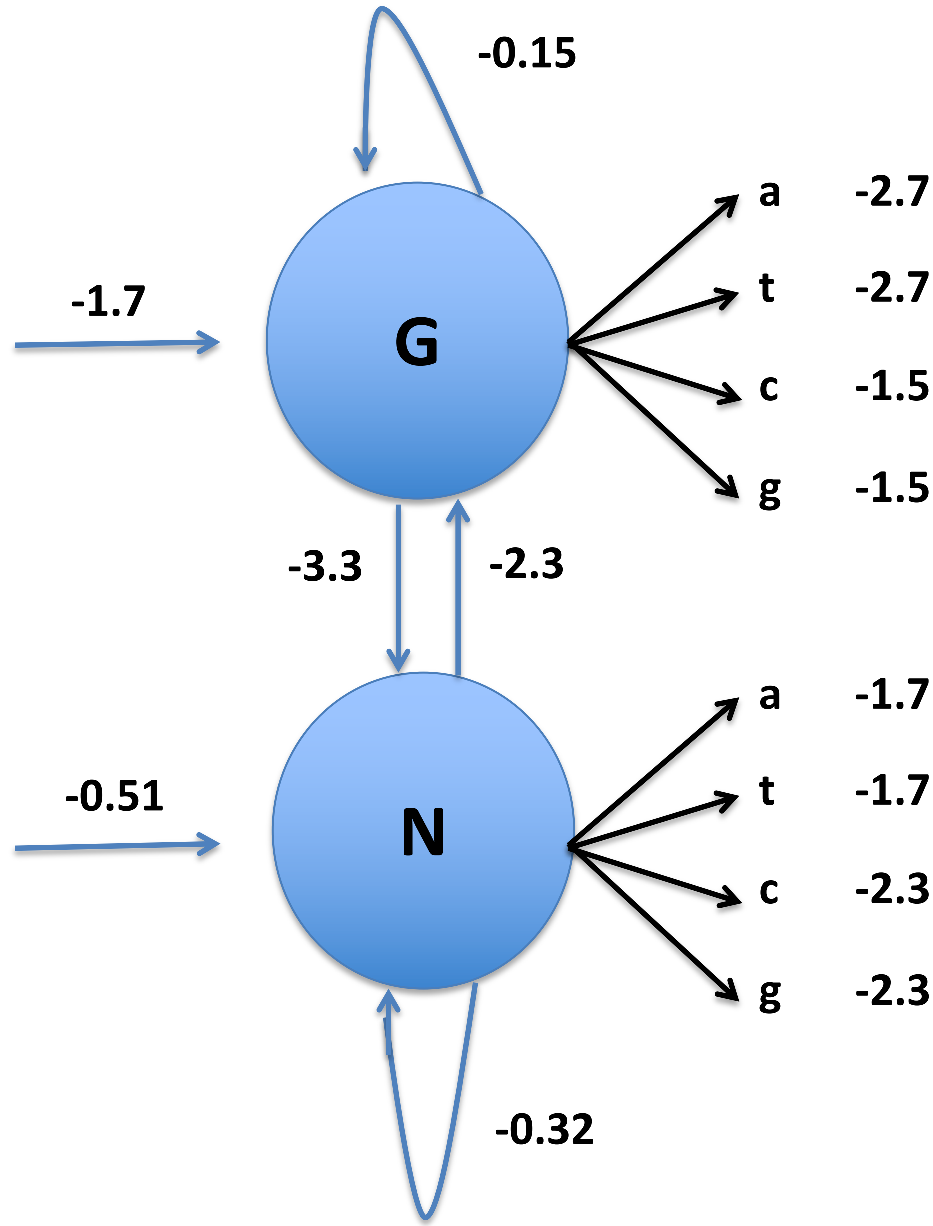


Part IV. Consider the HMM on the right, which models gene regions (G) and non-gene regions (N) in the genome, based on the fact that genes have higher GC content (guanine and cytosine nucleotides) than non-gene regions. Suppose the following sequence is observed: ***aaggc***

Note: In the questions below, you must show all your work to receive credit. <u>Do not round any of your answers.</u>

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

Use the dynamic programming matrix on the next page and answer the questions based on this HMM.



All probabilities are on the log2 scale

Part IV (do not round any answers)

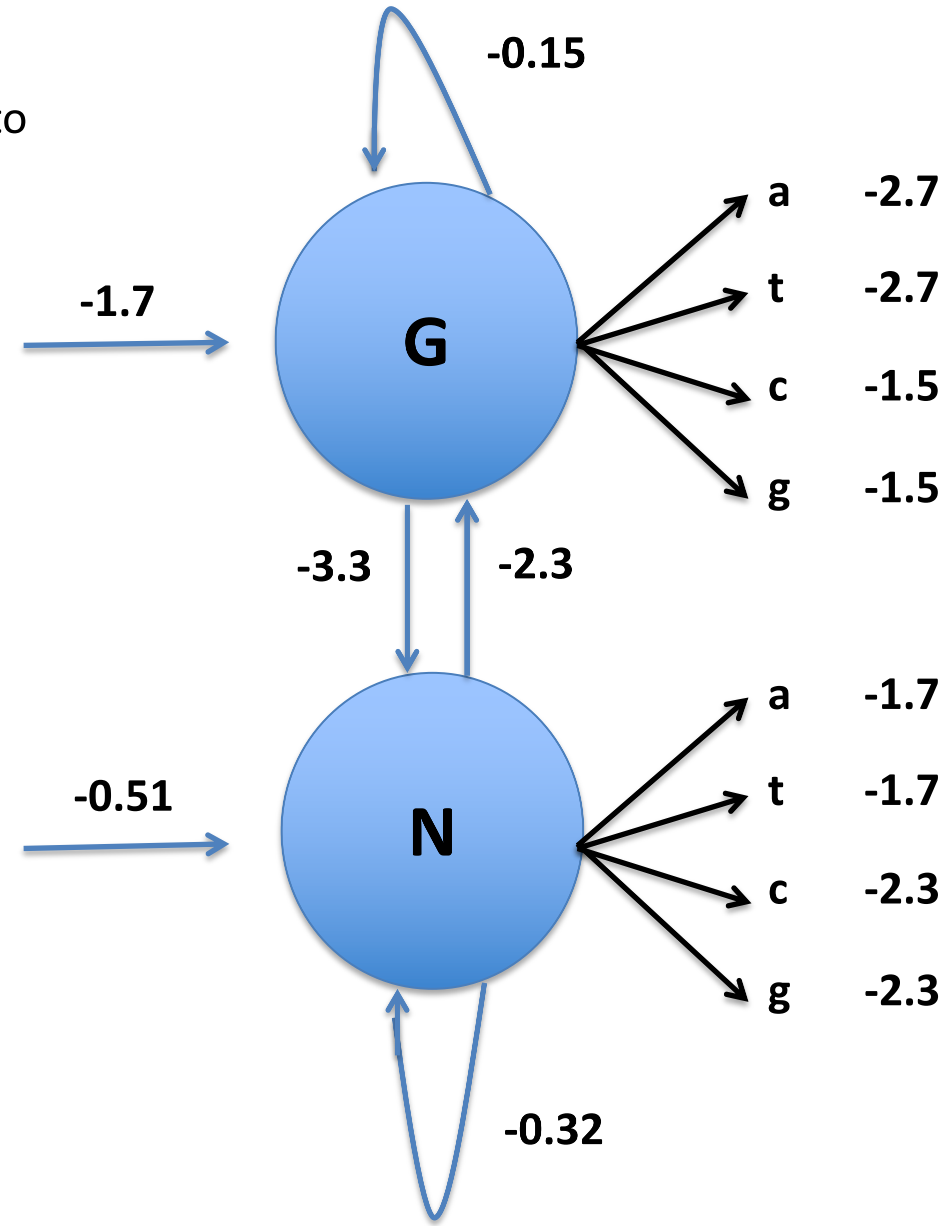
	a	a	g	g	c
Gene (G)	-4.4	-7.21	-8.03		
Non-Gene (N)	-2.21	-4.23	-6.85		

1. What is the optimal 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.
3. What is the optimal gene structure for the nucleotide sequence *aaggc*?
4. The probability that the optimal gene structure produced the sequence *aaggc* is proportional to what value?

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?

3. How does the Viterbi algorithm compare to a "brute force" approach that would require finding the probability of every possible state sequence?