# 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 5% of the time.

- 1. Specify the following probabilities:
    - P(Br) = probability a randomly selected female has invasive breast cancer
    - P(+|Br) = the probability that a female with invasive breast cancer will have a positive mammogram
    - P(+| healthy) = the probability that a female without invasive breast cancer will have a positive mammogram

- 2. Calculate P(+) = the probability that a randomly selected female would have a positive mammogram

- 3. Calculate P(Br|+) = the probability that a female with a positive mammogram has breast cancer. (This is the probability that a female who tests positivate really has breast cancer). Based on this result, is it likely that an individual with a positive mammogram has breast cancer? Note: this question provides insight into why the U.S. Preventive Services Task Force (USPSTF) advises against routine mammogram screening until women are 50 years old. More information: http://fivethirtyeight.com/features/science-wont-settle-the-mammogram-debate/

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: **THT**

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?

1.0

0.50

F

H   **0.50**

T   **0.50**

0.50

B

H   **0.90**

T   **0.10**

1.0
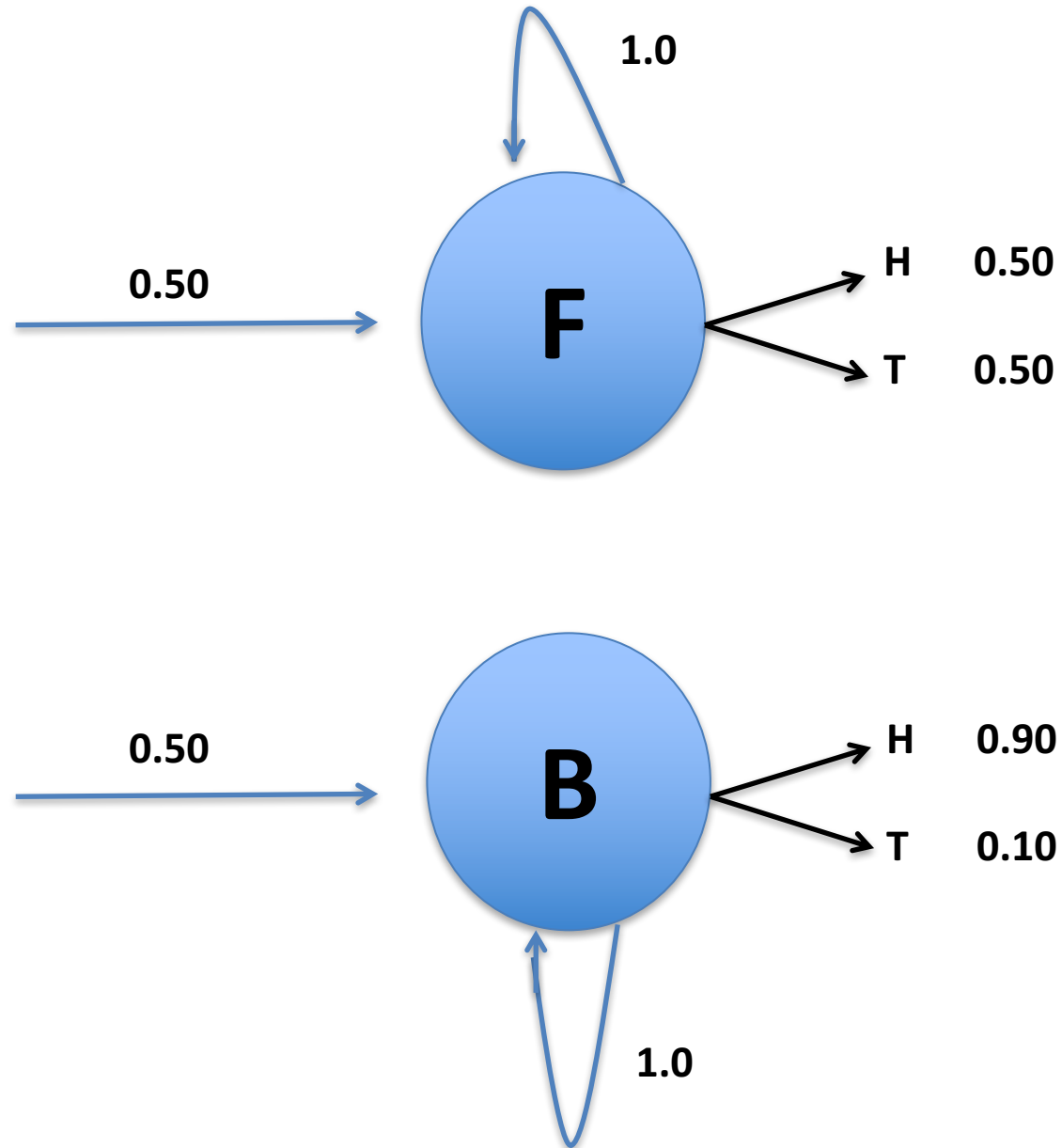
Part III. Consider the HMM on the right, which models gene regions (G) and non-gene regions (N) in the genome, based on the fact that genes have higher GC content (guanine and cytosine nucleotides) than non-gene regions. Suppose the following sequence is observed: *aaggc*
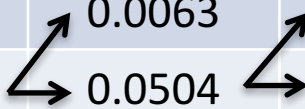
Note: In the questions below, you must show all your work to receive credit. <u>Do not round any of your answers.</u>

1. Given this observation of *aaggc*, show that the probability of the hidden state sequence NNGNN is proportional to 1.12896e-5 (0.0000112896)?
2. Use the dynamic programming matrix on the next page and answer the questions based on this HMM.

G:
a	0.15
t	0.15
c	0.35
g	0.35

N:
a	0.30
t	0.30
c	0.20
g	0.20

0.90
0.30
0.10   0.20
0.70
0.80
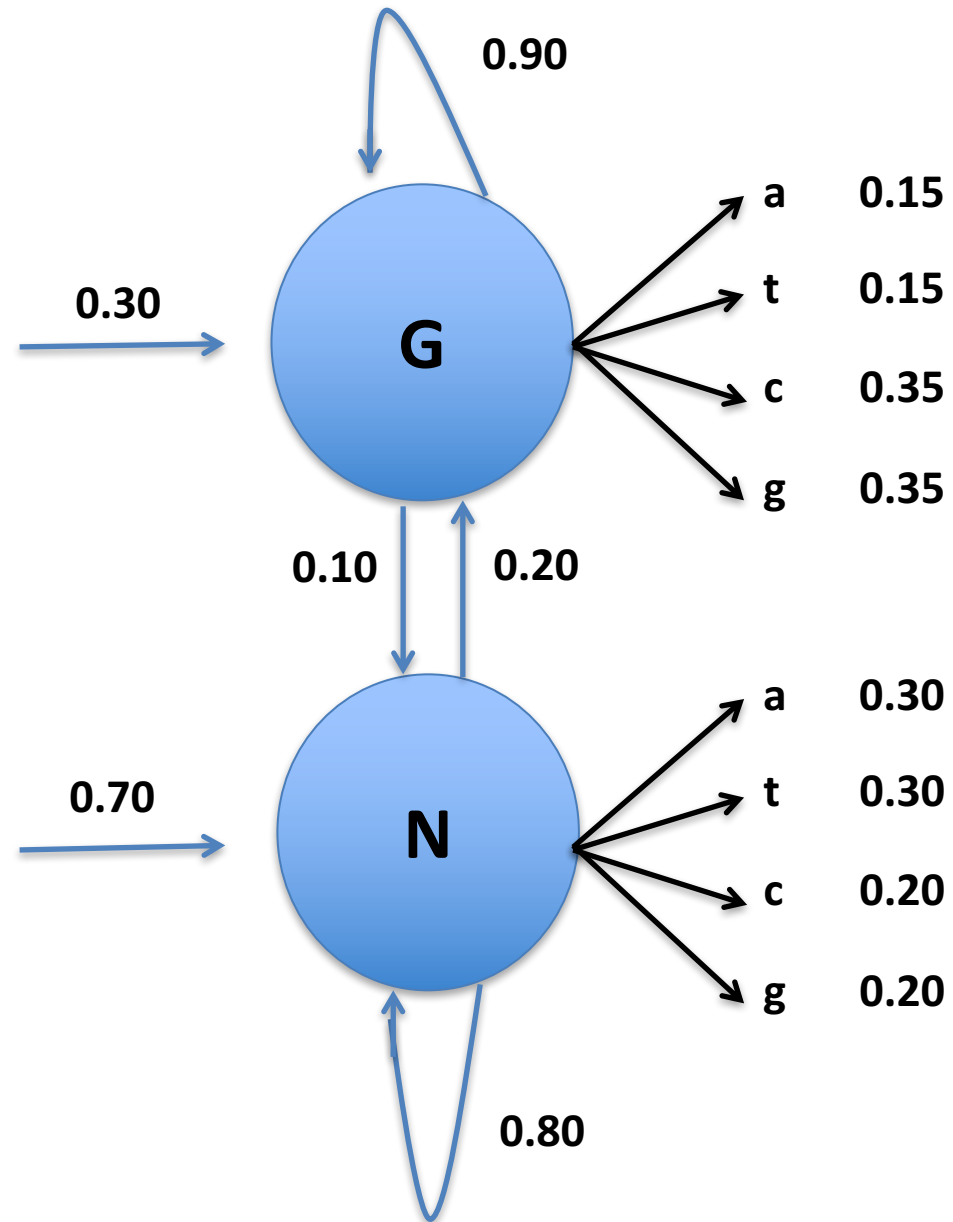
Part IV (do not round any answers)

	a	a	g	g	c
Gene (G)	0.045	0.0063	0.003528		
Non-Gene (N)	0.21	0.0504	0.008064		

1. What is the optimal gene structure for the dinucleotide sequence *aa*? The probability of that structure 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

0.90

1. Suppose that this same HMM was used to analyze a sequence 30 nucleotides long. What is the exact number of possible hidden state sequences? 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?

0.30

G

a   **0.15**
t   **0.15**
c   **0.35**
g   **0.35**

0.10   **0.20**

0.70

N

a   **0.30**
t   **0.30**
c   **0.20**
g   **0.20**

0.80