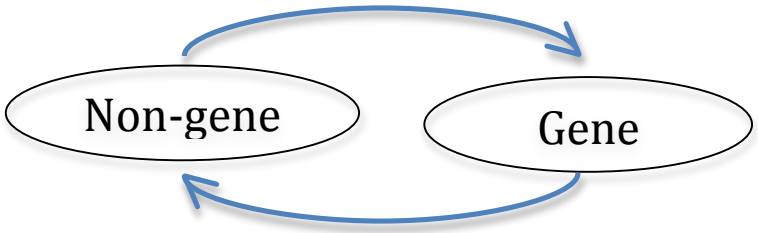
# COMP 462 / 561 - Homework #3
# Due on November 11 2019, 23:59 on MyCourses

**IMPORTANT:** Question 1 of this assignment requires a substantial amount of programming, and portions (c), (d), and (e) depend on you having a working program. **Do not leave this to the last minute!** Get started with the programming *early*, and get our help if you need.

**Question 1 (80 points).**
See code attached separately.

**Question 2. (20 points)**
If we think of an HMM as a machine to generate a random sequence of observations, the number of consecutive steps the path will remain at a given state follows a geometric distribution (https://en.wikipedia.org/wiki/Geometric_distribution ). This means, for example, that, in our gene-finding HMM, the length distribution of exons, introns, and intergenic regions will be assumed to be geometric. However, in reality, these regions have length distributions that can be far from geometric. Consider the very simple two-state gene finding HMM shown below. Assume that we have a desired gene length distribution, provided in the form of a discrete probability distribution for lengths ranging from 1 to 1000: Pr[length = k] = $p_k$.
Describe (in at most half a page) how you could modify the HMM below to produce a second HMM where the distribution over the duration of stay in the set of gene states is exactly the target length distribution. Note that the Gene state will probably need to be subdivided in several Gene sub-states. Describe not only the states of your HMM but also the transition probabilities.

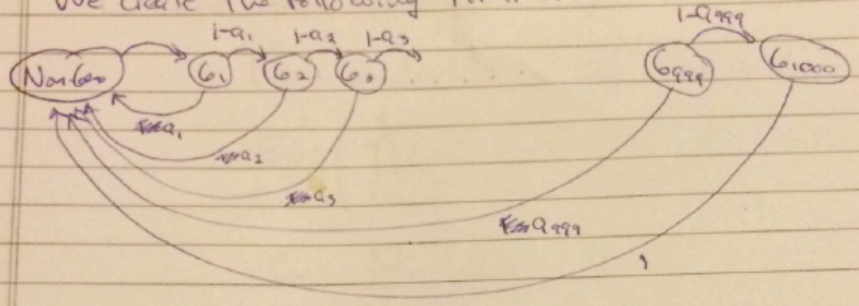



Good luck!

3. There are several solutions. I think the best is:

Assume $p_k = Pr[\text{length} = k]$ is given to us a target length distribution for genes

We create the following HMM:




How to choose $a_1 \ldots a_{999}$ so that the probability a gene of length $k$ is $p_k$?

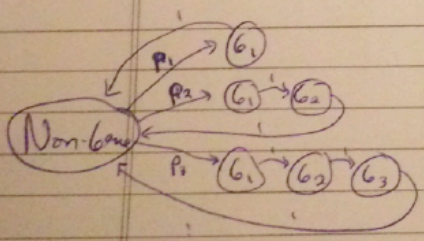
We want $a_1 = p_1$

$$(1-a_1)\cdot a_2 = p_2 \iff a_2 = p_2/(1-a_1) = p_2/(1-p_1)$$

$$(1-a_1)(1-a_2)\cdot a_3 = p_3 \iff a_3 = p_3/((1-a_1)(1-a_2))$$

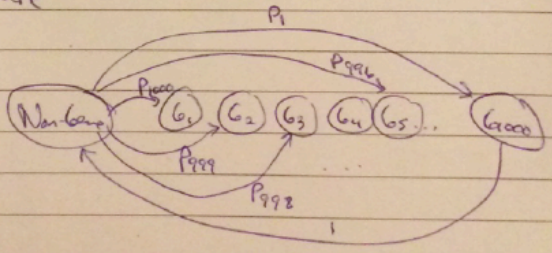
$$\vdots$$

Two alternate solutions are




(Less desirable because HMM gets very large)

or




(Less desirable because we can't assign position-specific emission probabilities)