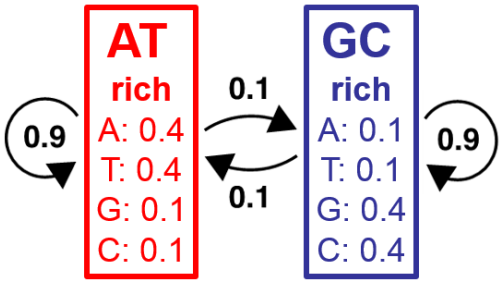
GS 373 Homework 7

Due May 23rd before 1:30 PM on Canvas

- (100 points): 4 bioinformatics questions (80 points), 1 programming assignment (20 points).
- Submit answers to the bioinformatics questions in a Microsoft Word document or PDF via Canvas. Your answers do not need to contain the text of the questions, but they need to be clearly labeled (e.g., 1a., 3b., etc.)
- Submit the programming assignment as a separate .py file onto Canvas. The script should be able to be directly run by Python.

Bioinformatics Questions (80 points)

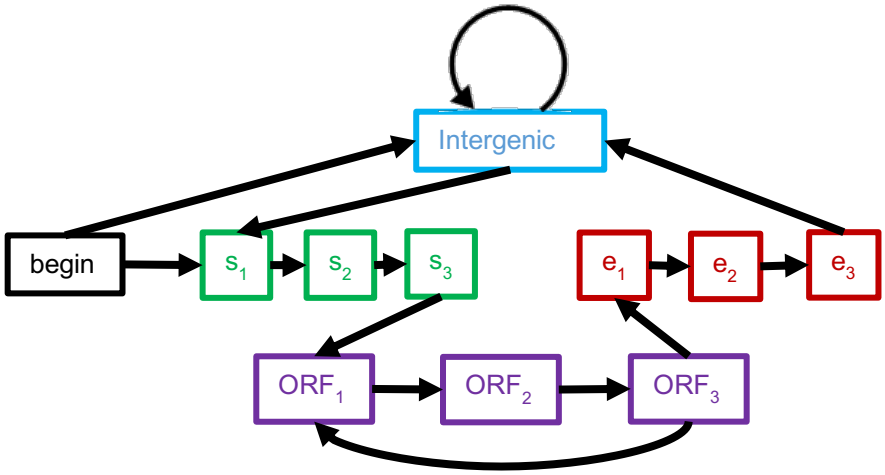
1. (30 points) **Consider the hidden Markov model that describes DNA sequences in two different states, AT rich and GC rich.** The numbers above the arrows denote transition probabilities, and the numbers within the boxes denote emission probabilities. Assume the initial probabilities of the two states are the same.



   a. (15 points) Based on this model, **calculate the joint probability of observing the sequence ACT with each of the following state paths**:
      i. The state path AT-rich, GC-rich, AT-rich
      ii. The state path AT-rich, AT-rich, AT-rich
      iii. The state path GC-rich, GC-rich, GC-rich

   b. (15 points) **Use the Viterbi algorithm to find the most likely state path for the observed sequence ACAT according to this model**. Show your work and be sure to label the state each base was emitted from.

2. (30 points) **The hidden Markov model below is a variation of the one proposed in class to identify open reading frames in DNA sequence, where the subscripts denote the position of a nucleotide within a codon.** "s" refers to the start codon, "e" refers to the stop codon, and "ORF" refers to the codons between the start and the end. s_1 refers to the first nucleotide in a start codon, ORF_2 refers to the 2nd nucleotide in a codon between the start and stop codons, etc.



a. (20 points) **What should the emission probabilities of each of the hidden states e_1, e_2, e_3 and s_1, s_2, s_3 be** if we assume that the three stop codons UAA, UAG, and UGA are equally likely in the organism of interest?

b. (10 points) Suppose you know that a specific nucleotide in an mRNA sequence is part of a coding region because it was shown experimentally to be directly bound by ribosomes. **Describe how you could use the HMM to infer the most likely reading frame for the mRNA sequence.**

3. (20 points) **Train the parameters of an HMM** with the two states A-rich and T-rich that describe sequences of A's and T's using the training sequence below, whose state sequence is also known.

Sequence	**A**	**T**	**T**	**T**	**T**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**T**	**T**	**T**	**T**
State	t	t	t	t	a	a	a	a	t	t	a	a	a	a	a	a	t	t	t

a. (10 points) **Compute the transition probabilities** for the states learned from the above sequence.
b. (10 points) **Compute the emission probabilities** for the states learned from the above sequence.

Programming question (20 points)

Implement a program that uses a Markov model to generate a string of As and Ts.

The main part of your program should perform the following steps:

1. read in the transition probabilities and store them in a workable data structure or structures
2. initiate a sequence
3. repeatedly add bases to it until the specified length is reached.

You will probably want to define a function that performs a single step in the Markov chain based on the previous state.

Your program should take two inputs as arguments using `sys.argv`:

1. The **length of the sequence** to be generated
2. An **input file containing transition probabilities** between A's and T's.

- This input file should be formatted as shown below. An example `transitions.txt` file is on the quiz section website. **Each line of the input file corresponds to a transition probability and contains three words separated by a space.** The **first is the name of the starting state**, the **second is the name of the ending state**, and the **third is the probability of transitioning** from the starting state to the ending state.

```
A A 0.7
A T 0.3
T A 0.2
T T 0.8
```

- Assume that there is an equal probability that the string with start with either an A or a T.

Below shows how the program should be used and expected output:

```
python homework7_skh.py 10 transitions.txt
```

```
AAATTAAAAT
```

Hints:

As a reminder, you can split a string using the string method split as follows:

```
>>> probs = "A T 0.9"
>>> probs.split(" ")
['A', 'T', '0.9']
```

As discussed in quiz section 7, the function `random()` returns a uniformly distributed random number between 0 and 1.

```
>>> import random
>>> random.random()
0.31762
```

Make sure you convert any numeric values that are read in as strings (arguments or from the file) into integers or floats before using them!