GS 373 Homework 2

Due April 19th before 1:30 PM on Canvas

- 100 points: 5 bioinformatics questions (80 points), 1 programming assignment (20 points)
- Submit answers to the bioinformatics questions in a Microsoft Word of 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 via Canvas. The script should be able to be directly run by Python.

Bioinformatics Questions (80 points)

1. (20 points, 3 parts) Complete the dynamic programming matrix and **calculate both the optimal Needleman-Wunsch global and Smith-Waterman local alignment** for the following pair of sequences, using the substitution matrix below and a linear gap penalty of -4. As with last week, please **include the arrows in your matrix** (either by constructing by hand and submitting a photo or by adding the arrows using a program like powerpoint) **as well as the best global and local alignments**.

substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

gaps = -4

CATGT and GGTT

a. Global

		G	G	T	T
	0	-4	-8	-12	-16
C	-4	-5	-9	-8	-12
A	-8	-4	-5	-9	-13
T	-12	-8	-9	5	1
G	-16	-2	2	1	0
T	-20	-6	-2	12	11

G	G	T	-	T
|	|	|		|
C	A	T	G	T

b. Local

		G	G	T	T
	0	0	0	0	0
C	0	0	0	0	0
A	0	0	0	0	0
T	0	0	0	10	10
G	0	10	10	6	6
T	0	6	6	20	16

G	T
|	|
G	T

c. Explain why the best global and local alignments are not identical.

Lots of possible explanations here. Local does not allow the score to go below zero, global must include all bases, local need not, etc.

2. (10 points) Given the completed dynamic programming matrix below**, write out all Smith-Waterman local alignments with the top score**.

		A	G	T	T
	0	0	0	0	0
G	0	0	10	6	2
T	0	0	6	20	16
A	0	10	6	16	15
G	0	6	20	12	11

A	G
|	|
A	G
G	T
|	|
G	T

3. (15 points, 3 parts) For each situation below, **state whether it would be more appropriate to perform a global or a local alignment and why**.
   a. Aligning 100-bp sequencing reads to a reference genome

Local alignment, a global alignment would require every base in the reference genome to be matched.

b. Identifying potential regions of homology between two distantly related bacterial proteins

Local alignment, only pieces of distantly related proteins would be expected to align.

c. Comparing an HIV gene sequence isolated from two different infected individuals

Global alignment, we would expect the HIV genome from different individuals to be mostly the same, with variants. (Accepted local with good reasoning)

4. (15 points) BLOSUM alignment scores. *Please show work for partial credit.*
   Consider the construction of a BLOSUM matrix given some gene blocks.
   a. The frequency of D in the gene block is 0.02, frequency of G is 0.03, and the half bit score of D-G is 4. **What is the frequency of D-G in the gene block? What is a probabilistic interpretation of the half bit score of 4 in this context?**

0.0048.

$4 = 2 \log_2(x / (2 * 0.02 * 0.03)$

2^(4/2) more likely to be seen in the blocks than expected due to chance

   b. **How does the half bit score change if the frequency of D increases to 0.03** (the frequency of G and D-G remains the same)? **How does the half bit score change if the frequency of D decreases to 0.01** (the frequency of G and D-G remains the same)? **What is a probabilistic interpretation of the *change* in the score** (not the score itself)?

When D increases, the frequency of D-G expected due to chance increases leading to a decrease in score. The opposite is true when D decreases.

   c. The half bit score of R-W is 0. **Given example values of the frequency of R, the frequency of W, and the frequency of R-W which could produce this score.**

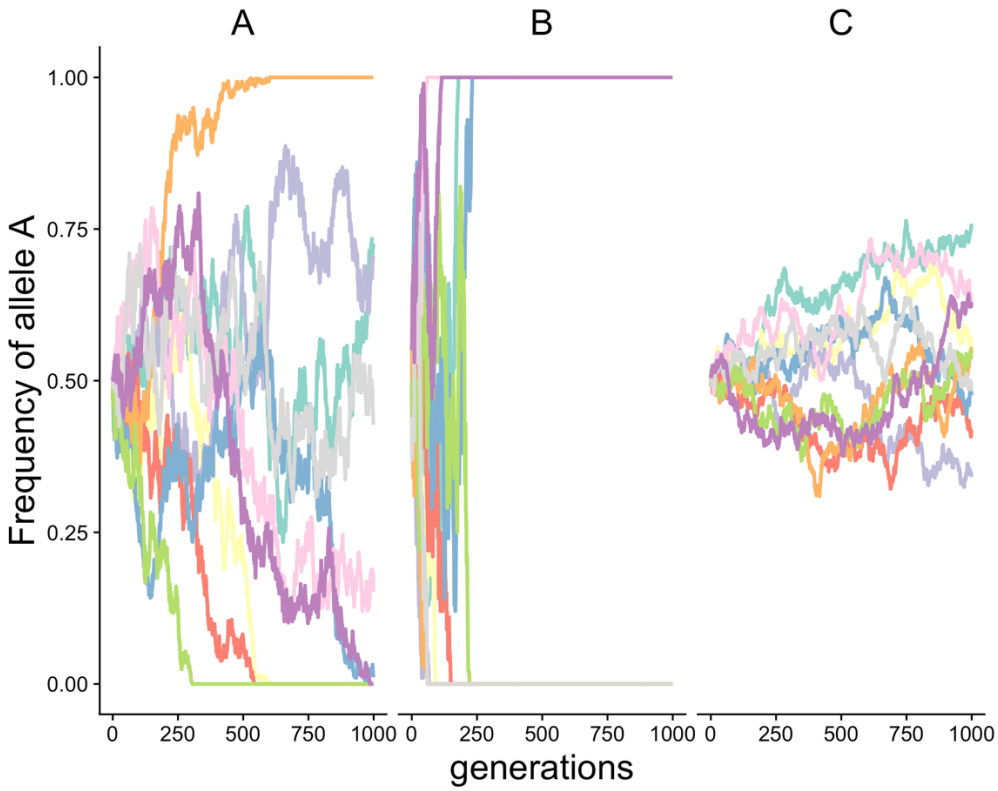
Many different answers. Anything that satisfies (2 * R * W) = R-W. ie 0.5, 0.5, 0.5

5. (20 points, 3 parts) Molecular evolution questions.
   a. One measure of evolutionary "distance" is the percent amino-acid identity (normalized hamming distance) of two sequences. For example, the sequences KAKLI and KWKLV have an amino-acid identity of 60% (3/5 amino acids match). Humans and pigs diverged about ~99 million years ago and their homologs have an average amino-acid identity of ~85%. Homologs from Influenza A Virus strains, which descended from the 1918 pandemic (~100 years ago), *also* have an average amino-acid identity of ~85%. **What could explain this difference?**

Many different answers. Generation time, mutation rate, population size, etc

   b. Below are three allele trajectories created via simulation. All three simulations share the same parameters *except* for population size. **Label the plots according to population size (1=largest population, 3=smallest population) and explain how you knew**.

A=2, B=3, C=1



c. Download and try running the script called `popsim.py`. Feel free to open the script and look at the code. This script simulates allele frequencies for 500 generations and then prints whether allele A fixed, allele a fixed, or neither fixed. You can change the population size, the starting frequency of A, or the fitness of either AA, Aa, or aa using the command line options `--pop`, `--freq`, `--fAA`, `--fAa`, and `--faa`, respectively. For example, the call `python popsim.py --pop 5000 --freq 0.3` sets the population to 5000 and the starting frequency of A to 0.3. **Run the script a few times and then post two screenshots**: one showing the results of a run with the default parameters and one with a run where either A or a fixed. Make sure the screenshot shows the full output. **Write a short, informal summary of your experiment** (~5 sentences). Are there differences between the runs? If so, what did you change (if anything)? If you made changes, what result did you expect when you made the change and did the results match your expectation?

Many different answers