BCH394P/364C Systems Biology/Bioinformatics **Marcotte**
Problem Set #1 **due before 10PM 2/2/2026**

Turn in all of your homework as a single pdf or text file submitted through Canvas, including work and programs for each problem, e.g., for # 1, turn in the program and the nucleotide percentages; for # 2, the program, program output and the percentages, etc.

A couple of Python problems

1. Write a short Python program to calculate the percentages of nucleotides in a DNA sequence. This can be the program you were given in the Python primer (it's perfectly fine to use the example program given in class) or a program of your own construction. Run it on the *H. influenza* genome and the *T. aquaticus* genome (get the nucleotide sequence files directly from the course web page). Turn in the nucleotide percentages as well as the absolute counts of the two genomes & your program.

2. Write a short Python program to calculate the percentages of all dinucleotides in a DNA sequence. Again, this can be based on the program example in the Python primer or of your own construction. (Notice that this is slightly trickier than it sounds at first glance, as dinucleotides can be overlapping. For example, "AAA" contains 2 "AA" dinucleotides, and "AAAA" has 3. This means that you can't simply use the Python string.count function to count dinucleotides, as Python counts non-overlapping instances, not overlapping instances. Since you want overlapping dinucleotides, you will have to try something else, such as the python string[counter:counter+2] command, as explained in the Rosalind homework assignment on strings.) Run your program on the *H. influenza* genome. Turn in the program and the dinucleotide percentages and absolute counts of the *H. influenza* genome that are output when you run the program.

3. Run your Python program from problem #2 on the genome of *T. aquaticus*. As before, turn in the observed dinucleotide percentages and absolute counts.

4. Calculate the dinucleotide percentages that you would expect for *H. influenza*, based upon the percentages of the single nucleotides that you calculated for *H. influenza*. Are the observed dinucleotide percentages of the *H. influenza* genome that you found in problem 2 consistent with what you expected for the dinucleotide percentages? If not, what might account for the difference? Turn in the expected dinucleotide percentages & your speculations.

5. Run your Python program from problem #2 on the 3 mystery gene DNA sequences that you can download from the course web page. Print out the dinucleotide percentages of each. Based on the observed dinucleotide percentages, determine which genome each of the 3 genes is taken from. Turn in each of the genes' dinucleotide percentages & your guesses for their identities.

Scale of biological data

6. Assume for simplicity that one nucleotide takes one byte of storage on a computer. Could I store my entire genome on my cell phone's 200 gigabyte SD card (1 GB is 10^9 bytes)? Despite the first human genome sequence costing a billion dollars, it is now technologically & economically possible to sequence a complete human genome for ~$500, and the price is still dropping. How much would it cost to sequence the genome of everyone in the US newly diagnosed with cancer in a given year? You can find up-to-date cancer statistics here: https://www.cancer.gov/about-cancer/understanding/statistics How much space would it take to store those genomes? Round your answer to the nearest petabyte (PB,

10^{15} bytes). For comparison, the UT TACC supercomputer file archive has a storage capacity of ~1 exabyte (1,000 PB).

7. How many times larger is the human genome sequence than the *E. coli* genome? The *E. coli* genome contains ~4,500 known genes, and the current estimate for the human genome is ~20,000-25,000 genes, or around 5 times more genes than in *E. coli*. Making the assumption that the genes of *E. coli* & human are about the same size (which isn't at all true), calculate the density of genes in the two genomes.

Lastly, some amino acid substitution matrix problems
Some exercises to familiarize you with the properties of protein sequences. Consult the BLOSUM50 substitution matrix in the class handout to answer these. The answers may vary depending on how you define each case, so be sure to explain your logic. (In other words, there is often more than one way to answer these—I'm interested in seeing your choice and in how you defend your answer.)

8. Which amino acid is most likely not to be substituted for by another? Explain how you came to this conclusion.

9. Which amino acids are most easily substituted for by others? Explain how you came to this conclusion.

10. What are the most disfavored amino acid substitutions? Explain how you came to this conclusion.

11. (Adapted from Exercise 2.1 in Durbin *et al*.):
Amino acids D, E, and H are charged; V, I, and L are hydrophobic (greasy).
What is the average BLOSUM50 score within the group of the 3 charged amino acids?
Within the 3 hydrophobic amino acids?
Between the 2 groups?
Suggest reasons for the pattern observed.