BIO337 Systems Biology/Bioinformatics **Marcotte**
Problem Set #1 **due before midnight 1/28/2014**

For each problem, turn in all of the products of your analysis in a single text file emailed to the TA (e.g., for # 1, turn in the frequencies; for # 2, turn in the program output and the frequencies; and so on.)

A couple of Python problems

1. Write a short Python program to calculate the frequencies of nucleotides in a DNA sequence. This can be the program from the Python primer or a program of your own construction. Run it on the *E. coli* genome and the *T. volcanium* genome (you can get the nucleotide sequence files from the BIO337 web page). Turn in the nucleotide frequencies of the two genomes (& your program, if it was your own design).

2. Write a short Python program to calculate the frequencies 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. Run it on the *E. coli* genome. Turn in the program and the dinucleotide frequencies of the *E. coli* genome that are output when you run the program.

3. Run your Python program from problem #2 on the genome of *T. volcanium*. Turn in the observed dinucleotide frequencies.

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

5. Run your Python program from problem #2 on the 3 mystery gene DNA sequences that you can download from the BIO337 web page. Print out the dinucleotide frequencies of each. Based on the observed dinucleotide frequencies, determine which genome each of the 3 genes is taken from. Turn in each of the genes' dinucleotide frequencies & your guesses.

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 an 8 Gb memory stick (8 gigabyte, or 8×10^9 bytes) such as the ones you can carry in your pocket)? On a 1 Tb (1 terabyte, 10^{12} bytes) hard drive, how many times would the human genome fit? In spite of the first human genome sequence costing a billion dollars, it is now technologically & economically possible to sequence a complete human genome for *almost* ~$1000. How much space would it take to store the genomes of everyone in Austin (assuming we were foolish enough to store the data without compressing it somehow)?

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.

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