Week 5: DNA Distance and Maximum Likelihood Assignment

Part 1: Write a program that:

- loads the sequences in data_for_jukes_cantor_exercise.fa
- determines the number of identical bases (S), not counting gaps, and the number of different bases (*D*), not counting gaps, in the sequence (be sure to ignore positions where one or both sequences have a gap or 'N')
- calculate *p,* the proportion of different bases
- calculate the estimate of the number of mutations per site, *K'*, using the formula from the lecture: $K' = (-3/4)Log(1-4p/3)$
- **write the counts for *S*, *D*, the value of *p*, and the value of *K'* to the results file**
- define a function that calculates the log likelihood using the formula from the lecture: $Log(L(K \mid D,S)) = D \times Log(pd(K)) + S \times Log(1-pd(K))$
    1. where *pd(K)* is the formula for the probability that two bases are different, for a given value of K, as given in the lecture: $Pd(K) = (3/4)(1-Exp(-4K/3))$
- with this function, use two methods to find the value of *K* that gives the highest value of the function, using the values of *D* and *S* calculated from your sequences.
    1. Brute force.
        - Start out at a very low value of *K* and calculate the likelihood
        - Write a loop that tries out a slightly higher value of *K* than was used the last time through the loop
            - Calculated the likelihood and compare it to that for the previous value of *K*
        - When the likelihood stops going up, the current value of *K* is the value that makes the likelihood the highest.
    2. Use scipy.optimize.minimize_scalar()
        - You will need to have the function return the negative of the likelihood so that the minimization will find the 'maximum'
- **Write the estimates obtained using these two methods to the results file.**

Part 2: Write a program that

- Reads the species names and the sequences from the fasta file Myotis_aligned.fa
- Calculates the Jukes-Cantor distance between all pairs of sequences
- Fill up a distance matrix with these values
- **Write the distance matrix to a file using the species names to head the rows and columns and with formatting that makes it readable**.

Turn in your two programs, and your two results files