Homework #6

(due Wednesay, April 17)

You can use the mutt_gamete_event.tsv (click on this link and download) data for this homework. In that tab-separated spreadsheet, I have cleaned the original data into a series of events. Each event is either a recombination event (code 'R' in the first column) or an end of the chromosome event (code 'E'). In the second column, we have the number of bases that separate the event from the previous event.

This encoding loses the info about which stretches of the DNA are Dingo or Labrador. But because that information did not affect the likelihood, it is easier to not deal with it. The chromosome identity is also lost.

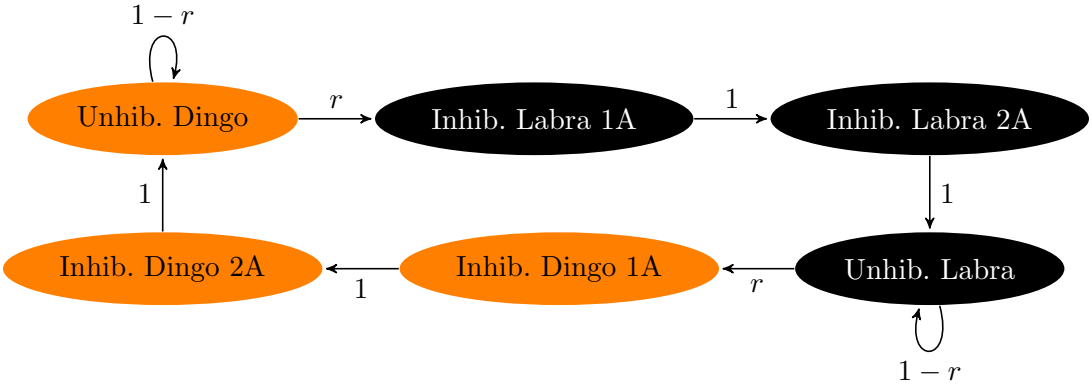
Email your code in with the homework. You can write your own code or use the python template or the R template provided.

#1. Find the MLE of the per-base recombination probability, r, using numerical methods. (Ideally this number will agree with what you calculated analytically in the previous homework).

#2. Use the 1.92 $\ln L$ drop rule to find a 95% confidence interval for r.

#3. Consider the following extension: Interference is the genetics name for "non-independence of recombination events." A very simple model of interference is that r describes the per-base recombination probability for "uninhibited" DNA, but a recombination event inhibits another recombination event from happening within a window around the point of recombination. Let's use w to denote the length of the window of inhibition. This zone of inhibition acts before and after each recombination event.

For example, if $w = 2$, you could imagine a state space like this:



where the "Unhib." states are outside of the window and can be subject to a random recombination trigger event. Note that the only real randomness is in the stretches of "Uninhibited" bases. The transitions in the inhibited states all occur with probability 1.

The start state for the chromosome would be in either the "Unhibited Dingo" or "Uninhibited Labrador" state with probability 0.5

The task will be to find the MLE of w and r, and do a LRT to see if this model of interference fits better than the model with no interference effects.

Fortunately, we don't have to draw a complicated state-space diagram for every different value of w. You just need to extend your likelihood model such that:

1. If 2 recombination events happen within w of each other, the likelihood is 0; and
2. The probability of no recombination at between any bases that are within w of a recombination event is 1.

This is a 2-parameter optimization problem (r and w must be jointly optimized). Also note that w is an integer, but numerical optimizers will send in floating point numbers for w. I would recommend just rounding the w parameter down to an integer in the likelihood. This will mean that the likelihood function is a step function in the w dimension. This can create optimization problems.

One alternative is to get a rough estimate of w using the 2-dimensional numerical optimizer, and then use for-loops to examine higher and lower values of w around the initial rough guess.

Another possibility helpful hint: you might be able to figure out the largest possible w for your data set and provide that as an initial bracketing constraint.