Homework 2

Instructions. As usual, you may work together on homework, but you must write up the solutions on your own. If you turn in an electronic copy, please provide it in pdf format. This homework is due by the beginning of class on Thursday, February 2nd.

Please read the updated lecture notes before attempting this homework.

There is a fastq-formatted NGS dataset with sequences from an uncultured archaeon on the website. We will use this dataset as a source of real sequences for hypothesis testing.

1. The 802nd sequence in the fastq file contains a long stretch of G's. Is such a long stretch of G's in this sequence unusual? You will address this question in the following parts.
   (a) We will be building a null sequence model H_0 to mimic the generation of this sequence. What structure, assumptions, or properties of H_0 would affect the chance of a run of G's? Why? And which of these structures, assumptions or properties, are most likely to be true in biological sequences?
   (b) Whatever H_0 we will ultimately choose below, what test statistic is sensitive to the truth of H_0 vs. the generic alternative H_1 where long runs are not unusual? What values of the test statistic indicate against H_0 and for H_1 (*i.e.* which values are extreme)?
   (c) Notice the question of interest is focused on *this* particular sequence. If you assume the sites of this sequence are iid, what exact test could you implement to test the significance of the run? Implement the test to compute the p-value (Important hint: You will want to use a Monte Carlo approximation!). What is the variance of your *estimated* p-value? (Reserve conclusion until part h.)
   (d) Continuing to make the iid assumption, what bootstrap procedure (sampling with replacement) could you use to test the significance of the run? Implement the test and estimate a p-value for the test. What is the variance of your *estimated* p-value?
   (e) The most complex parametric version of the iid model assumes sites are iid Categorical($\{A, C, G, T\}$) (or generalized Bernoulli or multinoulli, says Wikipedia) with pmf p_A, p_C, p_G, p_T. Find the MLEs for the four parameters (notice: there are three free parameters).
   (f) Is there any difference between the bootstrap model you repeatedly sampled from in Part d and this Categorical($\{A, C, G, T\}$) distribution? Which should you prefer for computing the p-value? Why? (Hint: Also see Part g.)
   (g) Another, less complex parametric iid model has two parameters, $p_R := p_A = p_G$ and $p_Y := p_C = p_T$ (here ":=" means "defined as"). Compute the p-value under this model. (Hint: You may prefer to estimate the p-value with Monte Carlo methods.)
   (h) What do you conclude? Is the run unusual? What is the difference (if any) between the tests? What did you discuss in a that these models ignore? Should you worry about the models used in these tests?

(i) Does it matter if the run of G's was identified by looking at the sequence? If so, could you modify the test so that you could trust the results?

2. In this question, we will use the rest of the data. The awful truth is that the run of G's in question 2 was selected by your very own instructor by looking at *all* the sequences. She didn't pick the longest run, but she picked a long run. (In my defense, I did it to make it interesting for you!) As a result, the p-values you computed in question 2 are definitely not to be trusted. (This does not answer to 2i, where you were to assume the 802nd was the only sequence examined.) In this question we will address a question without manipulating the data. Is the occurrence of homopolymer runs unusual in this set of sequences?

   (a) Implement a test that will address this question. Go through the usual steps: state H_0 and all of its assumptions, choose a test statistic, derive or estimate a sampling distribution for the test statistic assuming H_0, compute the p-value, and draw a conclusion. There are multiple valid solutions you could provide. I am looking for careful and reasoned solutions. I will also offer extra credit for particularly nice solutions.

   (b) Based only on the properties of the sequences you have uncovered and what you know about sequencers (or can easily google about them), what kind of NGS technology do you think produced these reads? Why? (You may, of course, verify your answer by looking up the accession number.)