Stat 610 Homework 2

Due Thursday, September 21, 11:59pm

Background

Scientists who do experimental evolution (https://en.wikipedia.org/wiki/Experimental_evolution) perform experiments in which they grow bacteria under different conditions and look for evidence of genetic adaptation of the bacteria to the environment.

Suppose you have collaborators who perform such experiments. They have a certain strain of bacteria, and they know the sequence of its genome. They grow clones (meaning all with the same genome) of that bacteria under two different conditions, and they are interested in whether the bacteria evolve in different ways under the different conditions. They perform 100 replicates under each condition, obtaining one mutated sequence for each replicate/condition combination for a total of 200 mutated sequences

The sequences collected at the end of the experiment can be found in jfukuyama.github.io/teaching/stat610/assignments/sequences.csv, and the germline sequences can be found in jfukuyama.github.io/teaching/stat610/assignments/germline.txt.

You are responsible for deciding whether evolution seems to proceed differently under the two conditions. Your first idea is simply to see whether the overall number of mutations differs between the two conditions. That is, if g represents the germline sequence, g_j represents the nucleotide in the germline sequence at position j, s represents one of the mutated sequences, and s_j represents the nucleotide at position j in s, then

$$n_{\text{mut}}(g, s) = \sum_{j=1}^{p} \mathbf{1}(g_j \neq s_j) \tag{1}$$

After some more thinking, you decide that the number of mutations is too crude of a measure, and you ask your collaborators whether they know anything else about what they would expect in this problem. They come back and tell you that they actually have a pretty good idea about the probabilities of each base mutating into each other base. Most of the time there are no mutations, but when there are the mutations tend to be between A and G bases or between T and C bases, and based on other data they have collected, they think that the following matrix is a pretty good estimate of the probabilities.

$$T = \begin{pmatrix} 0.93 & 0.05 & 0.01 & 0.01 \\ 0.05 & 0.93 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.93 & 0.05 \\ 0.01 & 0.01 & 0.05 & 0.93 \end{pmatrix} \tag{2}$$

In this matrix, T_{ij} represents the probability that a germline base i will mutate into base j by the end of the experiment. The order of the bases here is A, G, T, C, so for instance, the probability that germline base A mutates into a T is given by T_{13}.

Based on this, you decide to use a likelihood-based statistic to describe the mutational pattern in the two conditions. For each sequence, you compute the likelihood of obtaining the mutated sequence from the germline sequence under the model your collaborators gave you, that is, if g_j is the germline sequence at position j and s_j is nucleotide in position j in one of the mutated sequences s, then your likelihood-based measure is

$$l(g, s; T) = \sum_{j=1}^{p} \log T_{g_j, s_j} \tag{3}$$

Assignment

Your assignment is as follows:

1. Write a function to compute the number of mutations in each sequence (as in equation 1).

   Your function should take two arguments, the starting sequence and the ending sequence, and should return one number.

   For example, if the input sequence is AGTTC and the output sequence is TGTCC, your function should return 2 (the nucleotide at the first site changed from A to T, the nucleotide at the fourth site changed from T to C, and all the others stayed the same).

   In the function, you should check whether the sequences match each other in size and whether they are valid DNA sequences, that is, they only have values A, G, C, or T. If either of these are not true, the function should exit with an error.

2. Write a function to compute the measure of sequence divergence in equation 3.

   Your function should take as arguments a starting sequence, an ending sequence, and a 4×4 transition matrix.

   Your function should perform the same checks as the function in the previous part, plus check that the transition matrix is valid, that is, it is of the correct size, the elements of the transition matrix are all non-negative, and all of the rows sum to 1.

3. Download the data from the website (`jfukuyama.github.io/teaching/stat610/assignments/sequences.csv`), and read it in using `read.csv`.

4. Use the functions you defined in parts 1 and 2 to compute number of mutations and the sequence likelihoods for each sequence in the data you downloaded.

   You can use either a for loop or, if you have seen them before, a function from the apply family.

   Compute and report the mean and standard deviation for each of the four combinations of experimental condition and statistic.

5. If you would like, try asking chatGPT or similar to write these functions for you. Comment on the quality of the automatically generated code.

Submission parameters

Submit:

- Submit one `.R` file containing the functions you wrote. You can start with the template at `jfukuyama.github.io/teaching/stat610/hw2-template.R`, which has skeletons for the functions, the matrix T, and code to read in the germline and the sequences.
- One `.pdf` file containing your answers to the question 4.