IB120/201 - Lab 8

Modeling DNA Evolution

Due Date: March 20, 2020

University of California, Berkeley *GSI: Naveed Ziari*

In this lab, we will read two nucleotide seqeunces and make inferences about their evolution. In this pursuit, we will dig deeper into our understanding of maximum likelihood estimation.

Background

According to the Jukes and Cantor 1969 model, the probability that two nucleotides are identical to each other is:

$$\frac{1}{4} + \frac{3}{4}e^{-\,{}^{4\lambda}/_{3}}$$

and the probability that two nucleotides differ from each other is:

$$\frac{1}{4} - \frac{1}{4}e^{-\,{}^{4\lambda}/_{3}}$$

So the likelihood function, assuming independence among sites, if d out of S nucleotides differ between the two sequences is:

$$\mathcal{L}(\lambda) = \left(\frac{1}{4} + \frac{3}{4}e^{-\,{}^{4\lambda}/_{3}}\right)^{S-d} \left(\frac{1}{4} - \frac{1}{4}e^{-\,{}^{4\lambda}/_{3}}\right)^{d}$$

Maximum Likelihood Estimation

Maximum likelihood estimation is a method for calculating the parameters of a probability distribution through maximization of the likelihood function. In the coin toss example from the previous lab, the likelihood function was a binomial with a parameter θ that informs on the fairness of the coin. For the Jukes and Cantor model above, maximum likelihood estimation requires plotting $\mathcal{L}(\lambda)$ as a function of λ and finding its maximum point, hence the name *maximum likelihood estimation*. In essence, MLE aims to find parameters such that the observed data has the highest probability of occurring.

Questions

1. Read in the two sequences using the function provided and calculate the number of pairwise differences d.

2. Implement a likelihood function based on the Jukes & Cantor model. The likelihood function is parameterized in terms of λ which is the genetic distance. $\lambda$is proportional to the mutation rate and the amount of time the two species have been diverging from each other. The type of genetic distance is fundamental in many phylogenetic studies and other studies comparing DNA sequences.

3. Create an R function that calculates the logarithm of the likelihood. This function requires 3 arguments: S,d, and λ.

4. Use the `optim` function in R to optimize the this log-likelihood function for λ using the value of S and d, from the two sequences. Does this value match the one observed in the plot?

5. (6) Transitions (changes between A and G and between C and T) tends to occur at a higher rate than transversions (all other changes). Calculate the number of transition ds and the number of transversions dv from the two sequences. The Kimura two-parameter model, is specified in terms of two parameters $\alpha$and β. These two parameters can be interpreted as the rate of transitions and tranversions per unit time, respectively, multiplied by the total amount of time. The likelihood function under this model is:

$$\mathcal{L}(\lambda) = \left(\frac{1}{4} + \frac{1}{4}e^{-4\beta} + \frac{1}{2}e^{-2(\alpha+\beta)}\right)^{S-dv-ds} \left(\frac{1}{4} + \frac{1}{4}e^{-4\beta} - \frac{1}{2}e^{-2(\alpha+\beta)}\right)^{ds} \left(\frac{1}{4} - \frac{1}{4}e^{-4\beta}\right)^{dv}$$

   Make a function in R that calculates the logarithm of this likelihood. It should take 5 arguments: S, dv, ds, α, and β

6. Plot the log likelihood function for the values of S,dt, and dv you obtained from the two sequences as a function of α and β in a contour plot. Approximately where is the maximum likelihood estimate? *For contour plots, please use the* `plotly` *package, please refer to this link.*

7. Use the `optim` function in R to optimize this log-likelihood function for α and β using the value of S, dt, and dv from the two sequences. Does this value match the one observed in the plot?