Statistical Genetics, Spring 24

Homework exercise 1

Due date: 3 July 2024 in class

1. **Investigating mutation rates in mtDNA coding region**
   This problem uses the list of mutations per site in coding region of mtDNA from the paper of Behar et al. (2008), which we used in class to estimate & test the negative binomial model. It is available from: http://tau.ac.il/~saharon/StatGen2024/counts.txt.
   The second column is the location in the sequence, the third is the number of mutations counted.

   (a) Refit the negative binomial model we fit in class ($\alpha = 0.16775$) and test its goodness of fit via:

      i. A Kolmogorov-Smirnov test (why is this not an ideal situation for this test?)

      ii. A Pearson's chi-square test (what information is lost here?)

   (b) Investigate the relation between the mutation rate and the location along the sequence:

      i. Perform a Poisson regression with the location as a covariate, linearly or quadratically.

      ii. Is this model preferable to the simple Poisson model with no covariates?

      iii. Compare this model to the negative binomial model in terms of goodness of fit.

   (c) The coding region of mtDNA is densely populated with coding elements of known function. However, it still has several small "islands" of non-coding DNA (e.g., positions 3305–3306). The mapping of the sequence can be found at:
   http://www.mitomap.org/MITOMAP/GenomeLoci.
   Investigate whether the mutation rate at the identified "non-coding nucleotides" is faster than in the rest of the coding region.

   (d) *** Extra credit challenge: testing for dependence**
   Design and perform tests for dependence of mutation rates between neighboring sites; between sites in the same functional element; etc. Any significant finding will get a bonus grade, but insightful ideas and discussion may also get a bonus.

2. **Generalization of Poisson parity result to four states**

   (a) Assume the Jukes-Cantor (69) transition matrix:

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array}$$

   Prove that for $i, j \in \{A, C, G, T\}$:

$$P_{ij}(t) = \begin{cases} 1/4 + 3/4 \exp(-4\lambda t) & i = j \\ 1/4 - 1/4 \exp(-4\lambda t) & i \neq j \end{cases}$$

(b) *** Extra credit:** Generalize the result to the K80 transition matrix:

$$Q = \begin{array}{cc} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix} \end{array}$$

3. **Playing with the 5-parameter STR mutation model of Whittaker et al. (2003).**[1]

   (a) Implement this model, and draw an approximate sample from its "stationary" distribution as follows:

   - Start at a random STR length
   - Make steps according to the model probabilities
   - Draw a sample every 100,000 time units, until you have collected 1000 samples[2]

   Plot the empirical distribution of the sample, and calculate its mean and variance. Compare it to the empirical distribution in Figure 4 of Whittaker et al. (2003).

   (b) Investigate (theoretically or empirically) under what conditions a stationary distribution exists. Specifically, assume $\lambda = 1.06$ is fixed and address what conditions on $\hat{\alpha}_u, \hat{\alpha}_d, \hat{\gamma}_u, \hat{\gamma}_d$ are required for existence of a finite stationary distribution.

[1] The paper is available from the class website.

[2] Note that most steps will include no mutation, and can be skipped in a smart implementation, for example using the exponential waiting time and drawing the time of the next move.