Exercise 2

This is a practical session that will take place on March 10, 2022. **Bring your laptop to class with the seqinr package installed and the aedesaegypti.fasta file downloaded**. The genome of the mosquito Aedes aegypti can be found in the file aedesaegypti.fasta. This is a mosquito that can spread dengue fever, chikungunya, Zika fever, Mayaro and yellow fever viruses, and other disease agents.

1. Do a sliding window analysis of the GC content, that is, to study the variation in GC content within the genome sequence:
   (a) calculate the GC content of chunks with length 200 and length 1000 (window sizes=200 and 1000)
   (b) find the maximum GC content for each window size and plot the GC content around the point ($\pm$ 1000) where the maximum is reached.
2. Fit the genome sequence to a Multinomial and to a Markov chain model. Estimate its corresponding probabilities and transition probability matrix. Compute also the BIC and decide which model is better.
3. Consider again sliding windows of length 50 and calculate the GC content and the presence/absence of the trinucleotid "aaa". Is there any relationship between the presence of "aaa" and the GC content? What is the probability of "aaa" for a chunk with a GC content of 0.51 ? Plot the estimated probability of "aaa" against the GC content.
4. Consider again sliding windows of length 50 and calculate the GC content and the counts of the trinucleotid "aaa". Is there any relationship between the mean of the counts of "aaa" and the GC content? What is the predicted mean of the counts of "aaa" for a chunk with a GC content of 0.4 ? Plot the estimated mean of the counts of "aaa" against the GC content.

You must work in pairs (two-person groups), and the couple must be different from the one used in Exercise 1. Deadline: 27/03.