# 6. Graded assignment number 1 - population genetics and phylogeny

AUTHOR
EPFL - SV - BIO-463

PUBLISHED
March 25, 2025

This problem set will be graded and count for **25% of your final grade**. You can discuss with TAs and with fellow students about the problem set, but in the end you should hand in a **personal** solution. Detected plagiarism will result in a reduction of your grade.

The expected language is R, as in all the class BIO-463. Below, some R functions and libraries are recommended.

Please hand in your solution, which should contain both the R code and the accompanying explanations and answers in the same file, in **two versions**: - One should be your source file, in **.Rmd** or **.qmd** format; - The other one should be the **html** file directly deriving from your source file.

The two exercises are fully independent from each other. Each of them will be allocated the same number of points.

Some questions are broken down into itemized sub-questions. The goal is to help you address all points. Some sub-questions only require a short answer, but please make sure you answer each of them.

Please hand in your solution on **Moodle** by **Friday, March 28 at 11:59pm**.

## Exercise 1 - Evolution of HIV

Here, we will analyze the evolution of HIV sequences using the *ape: Analyses of Phylogenetics and Evolution* package and the *adegenet* package.

The data file "sequences1.fasta" contains sequences extracted from HIV genomes that were collected from one single patient, who was followed in time. The first sequence is an exception: it is not from the same patient - it is a reference sequence that was collected much earlier. The file "annot1.csv" contains annotations for these sequences. In this annotation file, the number of days after seroconversion will be of particular interest to us. Seroconversion marks the start of the development of specific antibodies in the blood serum as a result of infection.

1. Load the sequences using *read.dna*. Load the annotations using *read.csv*, and inspect them, e.g. by displaying the first few rows of annotations.

2. In this question, we will build a phylogenetic tree for these HIV sequences:

- Calculate all pairwise distances between sequences using *dist.dna* under the Jukes-Cantor model.
- Next, infer a neighbor joining (NJ) tree from this list of distances using *bionj*.
- Use the first sequence (reference) to root the tree (i.e. as outgroup) using *root*, and then use *ladderize* to transform the resulting tree.
- Plot the tree using *plot*, annotating each leaf of the tree with the number of days after seroconversion that the sequence was collected at (using column "DaysFromSeroconversion" in the table of annotations), and representing this number of days visually by a color gradient as well as by a tip label for each leaf.

*Clue for question 2:* for visualization, you can tune figure height and width, e.g. by including the following two lines at the beginning of your code: "#| fig-height: 20" and "#| fig-width: 12".

3. Comment on the tree obtained:

- How are sequences collected at recent time points positioned on the tree in general?
- What does this reflect?
- How many exceptions are there to this trend, and when were they collected (in days after seroconversion)?
- What can we conclude about these exceptions?

The data file "sequences2.fasta" contains sequences extracted from HIV genomes that were collected from another single patient, who was also followed in time. Again, the first sequence is an exception: it is not from the same patient - it is a reference sequence that was collected much earlier (the same as before). The file "annot2.csv" contains annotations for these sequences.

4. Perform the same analysis as in question 2 on this data.

*Clue for question 4:* for visualization, you can tune figure height and width, e.g. by including the following two lines at the beginning of your code: "#| fig-height: 12" and "#| fig-width: 12".

5. Comment on the tree obtained:

- How many sequences collected more than 100 days after seroconversion appear on the tree?
- Where are they positioned on the tree?
- Is this surprising in light of the previous analysis (questions 2-3)?
- Propose a hypothesis regarding the difference between these two patients.

6. In this question, we will focus on sequences that were collected at seroconversion.

- Extract all the sequences from this second patient that were collected at exactly 0 day after seroconversion, and form a smaller list of sequences with these sequences.
- Extract their annotations too, in a small annotation table.
- Calculate all pairwise distances between this subset of sequences using *dist.dna* under the Jukes-Cantor model.
- Plot the histogram of these distances.
- Calculate their mean and their maximum value.

7. For the flu virus, the mean Jukes-Cantor distance between all sequences collected across the world in a given year is 0.008 and the maximum Jukes-Cantor distance between sequences in a given year is 0.04.

- Compare these values to those observed in question 6, in orders of magnitude.
- What does this tell us about HIV evolution?

*Note for question 7:* for HIV, seroconversion takes about one month after infection, and the vast majority of HIV-1 infections are initiated by a single, genetically homogeneous founder virus variant.

## Exercise 2 - Simulating experimental evolution with serial passage and mutations

In this exercise, we will analyze the evolution of a haploid and asexual population in the long-term evolution experiment (LTEE), which is an evolution experiment on *Escherichia coli* bacteria, performed with serial passages. For this, let us consider a model with serial dilutions such that the population of initial size $K$ (bottleneck) undergoes deterministic exponential growth for a time $t$ and then $K$ individuals are selected randomly from the grown population to form the next bottleneck (serial transfer), and so on. We assume that there are two types of individuals, wild-types with fitness 1 and mutants with fitness $1+s$. These fitnesses represent deterministic exponential growth rates. Here, for simplicity, we will assume that all mutants have the same value of $s$.

1. In this first question, we neglect all mutations, and thus we assume that there are only wild-type individuals in the population. In the LTEE, serial transfer is performed everyday. Population size is multiplied by 100 at the end of the day, compared to the initial bottleneck at the beginning of the day. Recall that we assume for simplicity that growth is exponential. What is the value of $t$ that matches the LTEE?

2. In this question, we will consider that some mutants may be present initially, but that no new mutation may happen during the growth phase.

- Write the mutant fraction $x'$ after growth as a function of the mutant fraction $x$ before growth and of the parameters of the system.
- Assuming that $s = 0.01$ and $x = 0.01$, taking $t = 5$, and taking $K = 5 \times 10^6$ for the bottleneck size (matching the LTEE), compute the value of $x'$.
- Same question if all values are the same except that $s = 0.5$.
- Compare the results obtained in these two cases, and comment.

3. To take into account the new mutations that may happen during the growth phase, we will use the function *rflan* from the R package *flan* that allows to sample the number of mutants that descend from a given wild-type cell after a certain number of generations. Specifically, *sampled_data=rflan(N, mutations = nmu, fitness = 1/(1+s), mfn = n_grow)* provides N realizations of the number of mutants that arise from growth starting from one cell and reaching n_grow cells, where nmu mutation events happen on average and mutants have fitness 1+s. The numbers of mutants sampled are stored in *sampled_data$mc*.

- Use *rflan* to sample N=1000 times the final number of mutants after growth with nmu=1, s=0.01 and n_grow=100.
- Calculate the mean and the variance of the final number of mutants over these N=1000 replicate realizations.
- How does the variance compare to the mean?
- What causes this?

4. In this question, we will consider a population that starts with $K = 5 \times 10^6$ wild-type individuals (to match the LTEE). The population undergoes exponential growth for one day as explained in question 1. We assume that all mutants have $s = 0.01$, and that one mutation event happens on average during the whole growth phase (nmu=1).

- Simulate 10 replicates of the outcome of this growth phase (with the same initial conditions), using *rflan*. Note: this takes a bit of computational time.
- Calculate the final fraction of mutants after growth in each case.
- Comment on the variability of this fraction across replicates, in light of the result of question 3.

5. In this question, we will consider a given dilution step, and we will call $k$ the number of mutants that are sampled to form the next bottleneck, which should comprise $K$ individuals.

- What is the name of the probability distribution that $k$ follows?
- Write the formula for the probability $P(k)$ to obtain a given value of $k$. Note: here, if possible, we recommend using LaTeX for presentation quality.

6. In this question, we will sample the number $k$ of mutants that exist at the next bottleneck, assuming that $s = 0.01$, $K = 5 \times 10^6$ and $x = 0.01$, and taking $t = 5$.

- Sample $N = 1000$ different values of $k$.
- Compute their mean and standard deviation.
- Plot the histogram of the values obtained.

7. In this question, we will simulate the serial passage model. We will neglect new mutations that may happen during the process. Thus, the only mutants are those present initially, and we will assume that they have an initial fraction $x = 0.01$.

- Simulate the serial passage model described above with $s = 0.01$, $K = 5 \times 10^6$ and $x = 0.01$, and taking $t = 5$, for 300 bottlenecks.
- Plot the fraction of mutants in the population versus the number of generations in 10 different realizations on the same plot.
- What are the characteristics and the long-term outcomes of these trajectories?
- What is this due to?