Evolution & phylogenetic reconstruction

Contents

In this computer lab you will experiment with simulated evolution and phylogenetic reconstruction: the reconstruction of the evolutionary history of species, simulated genes, or — in the end — language and music. We will work on this computer lab during several lab sessions throughout the course, and indicate which sections you are expected to finish in every lab session. If you don't manage to finish those sections, you can work on them at home, and ask questions on Canvas/by e-mail or during the next lab.

Last updated on August 30, 2019. Written by Bastiaan van der Weij and Dieuwke Hupkes in 2016. Updated by Peter Dekker (2017), Bas Cornelissen (2017, 2018, 2019) and Marianne de Heer Kloots (2018)

1 Getting started

1.1 R

We assume you have some basic understanding of programming and R. You will not have to write a lot of code yourself, but you should be able to read and run the scripts that we provide you with. If you have never seen any R code before, it might be useful for you to work through a tutorial, such as the first three pages of this tutorial by Bart de Boer. Of course, you can always ask us for help if you are stuck on a particular piece of code!

1.2 RStudio

We strongly recommend you use a particular environment for running your R code, called RStudio. If you haven't installed RStudio yet, you should do so by finding the appropriate download of RStudio Desktop (Open Source License) for your operating system on rstudio.com and following the instructions until you can open the program.

1.3 Console & Scripts

When you open RStudio, you will find your screen divided into a few different parts. The most important ones that allow you to interact with R are on the left. The bottom left panel, called the *Console*, is where you can enter commands directly, and where the results of your command will be shown immediately after pressing Enter. When R is ready to accept commands, the *Console* shows a > prompt. In the *Console*, you can retrieve commands that you entered previously by using the up arrow key. However, when you close RStudio, all the commands you entered in the *Console* will be forgotten. When you are using a sequence of multiple commands and/or would like to run your code more often than once, it is better to save them in a *Script* (a plain text file that contains your code). The top left panel in your RStudio window is the *Script editor*. Here you can enter commands that will not disappear after you run them (you can run a selected line of code from the *Script editor* by pressing Ctrl-Enter). The tab at the top of the panel will show the name of your script. R script filenames should end in `.R`. For the commands given in the labs of this course, we will always tell you if you should enter them in the *Console* or save them to your *Script*.

1.4 Working directory

To run a full script in R, you type `source("scriptname")` in the *Console*. To be able to run a script, it should be available to R. In RStudio, the easiest way to do this is to put them both in the same folder and set this folder as the *working directory*. In the labs for this course, you should always set your working directory to the folder with materials we provide for that lab (for this lab, that folder is

`lab-evolution-reconstruction`). Look at this website to find out how to set the working directory for your version of RStudio.

2 Simulated Evolution

The goals of this section are:

- to better understand the concepts of genotype, genotype space, fitness, fitness landscape, selection, mutation, selection-mutation balance, frequency dependent selection;
- to see how these concepts can be formalised in a computer program;
- to appreciate both the power and the limits of natural selection.

You need the following R files in the folder: and `auxiliary_functions.R` and `simulated_evolution.R`.

QUESTION 1

- Start RStudio and set the *working directory* to your locally saved `lab-evolution-reconstruction` folder.
- Open the provided `answers_simulated_evolution.R` file in the *Script editor* (click *File* > *Open File...* in RStudio). This will be the script to add your code to in this section of the lab.
- Install the package `stringr` by typing `install.packages("stringr")` into the *Console*.
- Load the library `stringr` by typing `library(stringr)` in your *Script*.

2.1 Creating a population

In the first part of this computer lab, we will use R to simulate the evolution of (DNA) strings under a particular fitness function. We will represent DNA strings (the *genotype*) as a sequence of the letters 'A', 'G', 'C' and 'U'. In R, you can generate a random sequence of 10 of these letters using the following command (try it out in the *Console*):

```
sample(c('A','G','C','U'), size=10, replace=TRUE)
```

To store the output in a variable (for instance `x`), you type:

```
x <- sample(c('A','G','C','U'), size=10, replace=TRUE)
```

You can then view the contents of a particular variable by simply typing its name (here, `x`) and pressing Enter. Let's experiment a bit with this.

QUESTION 2

- Using the commands you just learned, generate a few random sequences of length 10 containing the characters 'A', 'G', 'C' and 'U' to

confirm that it does what we want it to do.
- Generate a random sequence of length 50 containing the characters 'A', 'G', 'C' and 'U'.
- The set of all possible sequences is called the *genotype space*. How big is this space? I.e., how many genotype strings are possible with our representation?

Now, let's create a *population* of DNA strings. To do this, we will make 100 genotype strings. We will store our population in a matrix (the *population matrix*), where each member of our population is represented as a row of the matrix.

QUESTION 3

- Let's start with creating a matrix filled with zero's that we can later fill:[a] Add this line to your *Script*:

```
population <- matrix(rep(0, 100), 100, 50)
```

So what do we find in column 30 of our population matrix, and what does this number mean? - Fill your matrix by generating 100 population members in a for-loop and filling the matrix with them[b] (also add these lines to your *Script*):

```
population_size <- 100
for (i in 1:population_size) {
    population[i,] <- sample(c('A','G','C','U'),
      size=50, replace=TRUE)
}
```

[a] The command `matrix(x, height, width)` command transforms a vector `x` into a matrix with height `height` and width `width`.

[b] `x[i,]` accesses the ith row of the matrix `x`, which in our case thus corresponds with the ith member of our population

2.2 Defining a fitness function

Now we need to define a fitness function that computes the fitness of the individual members of our population. Imagine, for instance, that the string 'CAC' codes for some very useful aminoacid. The more CAC's in the genome, the higher the expected number of offspring. In our simulation of evolution, let's define the fitness as the number of times the substring 'CAC' appears in the genotype string (without overlap, so the string 'CACAC' contains *one* copy of 'CAC').

To keep track of the fitness of *all* members in our population (which are represented as rows in the population matrix), we create a *vector* where each element of the vector represents the fitness value for one member of the population.

QUESTION 4

- Generate an empty vector to store the fitness values, and call it `fitness` (add this to your *Script*):

```
fitness <- rep(0, population_size)
```

- Use a for-loop to fill the vector (created by the code above) with the fitness values (also add this to your *Script*):

```
# loop over population size
for (i in 1:population_size) {
  # generate string representation
  member <- paste(population[i,], collapse='')

  # compute fitness member
  fitness_member <- str_count(member, "CAC")

  # store in fitness vector
  fitness[i] <- fitness_member
}
```

(Note that R ignores everything that follows the character #. In programming terms, these texts are called *comments*.) What is the highest possible fitness a member of this population can have? - Compute the mean fitness of your population by using `mean(fitness)` in your *Console*. What is the average fitness of your population?

2.3 Implementing selection

Now we will generate the next generation. We assume that each member of the next generation inherits the genome of one of the members of the previous generation. The probability of inheriting each genome is proportional to the genome's fitness: a child is most likely to inherit the genome of the fittest member of the previous population. This simulates selection.

QUESTION 5

- Compute the average fitness of the population and store it in a variable (add this to your *Script*):

```
av_fitness <- mean(fitness)
```

- Generate 100 new children, using the built-in function `sample` (the same one we used before[a]) (also add this to your *Script*):

```
indices <- sample(100, size=100, replace=TRUE,
                  prob=fitness/sum(fitness))
new_population <- population[indices,]
```

- If one population member has fitness 10 and all the other population members have fitness 1, what is the probability that a child will inherit its genome from this one population member? What do you expect

to happen with the population?

[a] We first draw 100 random numbers between 1 and 100 (repetitions possible). If population member 2 has a very high fitness, it will have a very high chance of being drawn. Then we use the drawn numbers to create a new population of the members corresponding to the numbers.

QUESTION 6

- To simulate the evolution of the population, we want to repeat this process several times and plot the average fitness over time. If you like programming, you can try to do the implementation yourself, but we also provided a script called `simulated_evolution.R` that does the trick. You can run it by typing `source("simulated_evolution.R")` into the *Console*. The next bullet point contains some instructions for implementation, so if you use the script you can skip it.
- To repeat the previous process 100 times, you should create a for-loop that executes the previous bits of code 100 times, storing the fitness of every population in a vector. If you stored the fitness values in `av_fitness`, then you plot your results using

```
plot(seq(1,100,1), av_fitness, type="l", ann=FALSE)
# Add a title and label the axes
title(main="title", xlab="x label", ylab="y label")
```

- You will notice the fitness stops increasing quite early in the simulation. Why is this?

2.4 Evolution with mutation

In the previous simulation, we looked at selection *without* mutation. Let's now look at the case where every child's nucleotide has a probability μ to change into a random other nucleotide.

QUESTION 7

- If $\mu = 0.01$, what is the chance that no changes occur in a genome? What is the chance that no changes occur in an entire population? And if $\mu = 0.001$?
- Use the provided script `simulated_evolution.R` to do the same simulation, but with a mutation level $\mu = 0.001$. You can change the values of the parameters at the top of the script. Also adapt the length of the simulation to a number you think is suitable. After changing the parameters, save the file and run the script again.
- Now repeat the simulation with $\mu = 0.01$, and plot the fitness. This shows the mutation-selection balance.
- In the simulation with relatively high mutation rate, why does the fitness stop increasing at a slightly lower level?

3 Simulated evolution with ancestry

In the previous section we implemented computer simulations of evolution in R. In this section, we extend this simulation, and think about the patterns of genetic variation that the evolutionary process leaves in a population. We look at methods for *phylogenetic tree reconstruction*. These methods use the genetic variation in the most recent (current) generation to reconstruct the evolutionary history of a population or a set of species. We use a simple clustering algorithm[1] for phylogenetic tree reconstruction.

The goal of this section is to learn about the possibilities and difficulties that are involved in phylogenetic tree reconstruction.

You need these two R files in the `lab-evolution-reconstruction` folder: `auxiliary_functions.R` and `simulated_evolution_ancestry.R`. To plot the phylogenetic trees with R, you will also need to have the package `ggtree`. Install it by typing the following into the *Console*:

```
source("https://bioconductor.org/biocLite.R")
biocLite("ggtree")
```

3.1 Tracking ancestry

So far, we have simulated the evolution of strings of symbols, and looked at the effect of different fitness functions. Now we will repeat this simulation, but during the evolutionary process we will keep track of ancestry, so that we can reconstruct family trees of different individuals. We will start with a very simple simulation. The script `simulated_evolution_ancestry.R` runs the same simulation we saw in the previous part of this lab. This time, however, the script also generates a matrix called `parent_matrix` that specifies the parent of each member in each generation (where the parent is the individual of the previous generation whose genetic material was inherited). At the end of the simulation, a plot illustrating the development of both the average population fitness and the diversity of the population is generated.

QUESTION 8

- Change the parameters at the top of the file `simulated_evolution_ancestry.R`. Set both `population_size` and `simulation_length` to 10. What values do you expect on the y-axes of these plots? What do you think the curves of average population fitness and population diversity look like? At what point do you expect the curves to start and finish?
- Run the script by executing the following command in the *Console*:

[1] An algorithm is a description of a series of steps to do arrive at a certain end result or perform a calculation. Algorithms can for example be implemented by a computer program.

```
source("simulated_evolution_ancestry.R")
```

Are the results as you expected?

QUESTION 9

- Visualise the parent matrix by running the following from the *Console*:

```
print_parent_matrix(parent_matrix)
```

Where in this plot can you find the first generation? - Follow some paths up and down. Why do downward paths often end in dead ends, whereas upward paths always go all the way up?

QUESTION 10

Change the parameters back to their original settings:

```
population_size <- 100
simulation_length <- 1000
```

- Run the simulation again (this may take a while).

3.2 Plotting the family tree

Printing the parent matrix for such large simulations is not very helpful, because the network is too dense to properly visualise (you may try if you want). Rather than looking the parent matrix, we will use the parent matrix to reconstruct a *family tree* for only the last generation (that is, we only look at the members of previous generations whose offspring appears in the last generation.) We will then generate a visual representation of the tree.

QUESTION 11

- From the data you just generated, generate a family tree with the function `reconstruct_tree` and visualize it with the function `plot_tree`. Enter the following into the *Console*:

```
tree <- reconstruct_tree(parent_matrix)
plot_tree(tree)
```

- *If the* `plot_tree` *function for some reason does not work on your computer*, you can also use an online phylogenetic tree viewer. First generate a textual representation of the phylogenetic tree by typing `print_tree(tree)` in the *Console*. This prints a string with lots of brackets and numbers describing the tree in the so-called Newick format. Copy everything between the quotes. Go to icytree.org and click *File* > *Enter tree directly...*. Paste the textual representation you've just copied and click *Done*.

- As far as you can judge, how many generations ago did the LCA of the current population live?
- Which aspects of evolution leave traces that we can detect in the current generation and which aspects do not?

4 Phylogenetic reconstruction of biological data

In evolutionary research, the elaborate ancestry information represented by the parent matrix is usually not available. To reconstruct family trees we have to resort to different methods. Information about when species branched off ('speciated') can be deduced from genetic variation in the current population. For instance, horses are genetically more similar to donkeys than to, say, frogs. So, the last common ancestor of horses and frogs most likely lived much further in the past than the last common ancestor of horses and donkeys. In other words, the branch that would eventually evolve into frogs split off from the branch that would eventually evolve into horses earlier than the branch that would eventually evolve into donkeys. This type of analysis is called phylogenetic reconstruction. It's based on genetic similarity between members of the current generation, which we measure using a *distance measure*. There are R packages that can automatically perform this reconstruction. Lets start with installing these packages:

QUESTION 12

- Open the provided file `answers_phylogeny_biological.R` in the *Script editor* (click *File* > *Open File..* in RStudio). This will be the script to add your code to in this section of the lab. Install the packages `ape` and `phangorn` by typing in your *Console*:

```
install.packages("ape")
install.packages("phangorn")
```

Load them in your *Script* using:

```
library(ape)
library(phangorn)
```

The `phangorn` package comes with a dataset that contains real genetic data (i.e., RNA samples) from many different species. You can load this dataset by typing the following (add it to your *Script*):

```
data(Laurasiatherian)
```

To show a summary of the data you can type `str(Laurasiatherian)` in the *Console*. The data originates from the (now closed) Allan Wilson Centre in New Zealand. To find out more about this data, have a look at their website.

We will try to reconstruct a phylogenetic tree for these species. That is, we will try to reconstruct when different species branched off from each other, based only on genetic information of the current population (the last generation). The first

step is to measure 'genetic distance' between the genetic samples for each species. For simplicity, we assume that (1) all species ultimately originate from a single common ancestor (an uncontroversial assumption in evolutionary biology), and (2) that species have diverged genetically by picking up mutations at a roughly constant rate (a more problematic assumption). (Try and convince yourself that the phylogenetic tree reconstruction method we described requires the second assumption — and that this assumption is problematic when considering evolution in the real world.)

The distance between strings of DNA or RNA is typically measured by counting the number of mutations required to change one into the other. Because of the second assumption, the genetic distance between two species is proportional to the time that has passed since their last common ancestor.

QUESTION 13

- Select five species from the Laurasiatherian dataset (for instance three that you think are closely related and two that are more distantly related).
- Create a subset of the data containing just these five species using this line (add it to your *Script*):

```
mysubset <- subset(Laurasiatherian, subset=c(19,20,28,29,30)
```

The numbers correspond to the position of the species in the list printed by `str(Laurasiatherian)`, i.e. Platypus = 1, Possum = 3, etc.) You have to replace these numbers by the numbers corresponding to the species that you chose. - Verify that your subset contains the right species using the following from the *Console*:

```
str(mysubset)
```

- Compute the *pairwise distance* between all elements in the set using the function `dist.ml` and print it (add these lines to your *Script*):

```
distance_matrix <- dist.ml(mysubset)
print(distance_matrix)
```

The results are stored in a *distance matrix*. How can you read off the distance between two species from this matrix? Why are the numbers on the diagonal of this matrix zero? - Do the computed distances correspond to your intuitions about the selected species' relatedness? - Using pen and paper, or your favourite drawing software, reconstruct a phylogenetic tree that describes the evolutionary relations between your selected species. Use the principles described earlier. You shouldn't need to do any calculations.

5 Phylogenetic trees with hierarchical clustering

5.1 UPGMA

We can use a simple method called 'hierarchical clustering' to build the phylogenetic trees such as in the previous sections automatically. Hierarchical clustering can be done with the UPGMA algorithm. (If you need a refresher on how exactly it worked, look up the UPGMA chapter from the readings again!) In short, UPGMA follows these steps:

- Treat each datapoint (for example, a RNA sample of a species) as a separate "cluster" containing just one datapoint;
- Compute the distances between all clusters (using some distance measure; for example, genetic distance);
- Merge the two clusters that are nearest to each other into a new cluster;
- Repeat steps 2 and 3 until only all datapoints are in cluster.

5.2 Constructing a tree

To construct a phylogenetic tree, we can think of each merging of clusters as the joining of two branches. In the simplest version of this algorithm, we define 'distance' between a cluster A and a cluster B as the average distance between any datapoint in A and any datapoint in B — this is what UPGMA does. A slightly more complicated method, Ward's clustering, uses the square root of the average of the squared point-to-point distances.

QUESTION 14

Using the distances between species in `mysubset` (from the distance matrix you computed in the previous section), manually perform three cycles of the UPGMA algorithm with pen and paper.

The `phangorn` package we installed earlier provides pre-defined functions implementing different hierarchical clustering methods.

QUESTION 15

- Generate a phylogenetic tree for your subset and plot it using the following commands (add this to your `answers_phylogeny_biological.R` script):

```
tree <- upgma(distance_matrix, method='average')
plot(tree)
```

Is the tree the same as the one that you created before with pen and paper? - Create a tree for the entire dataset. Does it agree with your expectations? - *Optional:* Try different methods for computing the distance between clusters by changing the parameter `method` (options are, for instance, `'ward.D'`,

'single' and 'median'). Do you notice any changes in the resulting phylogenetic trees?

We will now investigate what happens if we perform phylogenetic analysis on the population resulting from our own simulated evolution. Remember that, since this is a simulation over which we have full control, we can reconstructed the *actual* phylogenetic tree using the information that we stored in the parent matrix.

QUESTION 16

- Open a new R script for this exercise: open the provided `answers_phylogeny_simulated.R` file in the script editor (click *File* > *Open File...* in RStudio). Add all the code in this exercise to your *Script*.
- Run the script `simulated_evolution_ancestry.R` again to generate a new population and parent matrix:

```
source("simulated_evolution_ancestry.R")
```

- Generate a distance matrix of the last generation from your simulation using the function `compute_distance_matrix`:

```
distance_matrix <- compute_distance_matrix(population)
```

- Reconstruct a phylogenetic tree with the `upgma` function (choose your preferred *method*, e.g. `'ward.D'`) and plot it:

```
tree <- upgma(distance_matrix, method='ward.D')
plot(tree, cex=0.3)
```

The parameter `cex` sets the font size of the plot, adjust it if the numbers are illegible.

- Now generate the *actual* family tree of the simulation by running

```
gold_standard_tree <- reconstruct_tree(parent_matrix)
plot_tree(gold_standard_tree)
```

How well does the reconstruction produced by the hierarchical clustering algorithm match the actual family tree?

- How can you explain the differences between the reconstructed and the actual family tree?

6 Phylogenetic reconstruction of languages

In the tutorial you have learned about features of language and music used in comparative research, and about variation across languages and musics. By studying some specific examples, you have seen that both languages and musical traditions are transmitted culturally, and are subject to a process of cultural evolution. In this final section of the computer lab, you will use the methods

that we saw for phylogenetic tree reconstruction in the previous sections to reconstruct the cultural evolutionary history of languages.

You will need the file `language_data.Rdata` from the `lab-evolution-reconstruction` folder, and you will also need to have the packages `ape` and `phangorn` installed.

QUESTION 17

Open the provided file `answers_phylogeny_language.R` in the script editor (click *File* > *Open File...* in RStudio). Add all the code in this section to your *Script*. We preprocessed the dataset for you so it can be loaded into R. To do this, type the following:

```
load('language_data.Rdata')
```

This will create an object called `mydata` containing the dataset. You can check the languages in the data by typing `names(mydata)` in the *Console*. Load the packages `ape` and `phangorn` using:

```
library(ape)
library(phangorn)
```

- Choose a subset of the list of languages. We will initially build a phylogenetic tree of this subset. Define your subset with the `subset` function. For instance, if you want to select language 40, 41, 42, 58 and 60 you type:

```
mysubset <- subset(mydata, c(40:42,58,60))
```

Create a distance matrix of your subset, by letting the computer count the number of feature values that differ between two languages ("hamming distance"):

```
distance_matrix <- dist.hamming(mysubset)
```

- Pick your favourite clustering algorithm and method and generate a tree, for instance:

```
tree <- upgma(distance_matrix, method='ward.D')
```

- Plot your tree:

```
plot(tree, use.edge.length=FALSE, cex=2)
```

- Do the same thing for the entire dataset (you might want to adapt the `cex` parameter, that sets the fontsize of the plot). Be aware of the influence the clustering algorithm and method for computing the distances between clusters can have.
- What are the nine main language families you can distinguish within the Indo-European family, and in which regions of the world are they spoken (before colonial times)?