Quantitative Life Sciences: from Infectious Diseases to Ecosystems

Jessy Duran Ramirez, Marco Labarile
jessyjomary.duranramirez@uzh.ch

14th December 2022

Molecular Epidemiology
Exercise sheet

Objectives:

- Understand that the choice of genetic distance measures influences the distance based reconstruction of a phylogenetic tree
- Being able to calculate the genetic distance of sequences using a model that accounts for evolution
- Understand one distance based algorithm to infer a phylogeny from sequence data
- Know that there are also more complex algorithms and being able to use one of them
- Apply the knowledge gained to “real-life” examples

R preparations:

```
install.packages("ape")
require(ape)
primates <- read.dna(file = "sequences_primates.fasta" , format = "fasta",
                     as.character = TRUE, as.matrix  = FALSE)
```

Exercise 1: Measure the similarity between sequences

Objective: Study different methods to measure the distance between sequences and to evaluate the impact of different distance measures on a phylogeny.

For this exercise consider the following example nucleotide sequences and calculate the pairwise distance according to the different models. Fill your result in to the tables provided.

orangutan	T C G C T G C C C G C A C A A T C A A T
gorilla	T C A C T G T C T A T A T A G C C A A C
human	T C G C T G C C T A T A C G G C T A A C
chimp	C C A C T A C C T A T A C G G C T A G C
mouse	A A C T G A T C T A T A C A T G T A T T

a) Write an R function to calculate the Hamming distance of the following (number of sites that differ between each pair of sequences):

	orangutan	gorilla	human	chimp	mouse
orangutan	____				
gorilla	____	____-			
human	____	____-	____		
chimp	____	____-	____	____	
mouse	____	____-	____	____	____

Hint: Your function could have the following form:

```
hamming_2seq <- function(seq1, seq2){
 ???
}

hamming_alignment <-  function(alignment){
n <-  length(names(alignment))
m <- matrix(nrow = ???, ncol = ???)
for(i in 1:n){
for(j in i:n){
m[j, i] <- m[i, j] <- hamming_2seq(???, ???)
}}
colnames(m) <- rownames(m) <- ???
return(m)
}

dist_ha <- hamming_alignment(primates)
print(dist_ha)
```

b) Implement the JC69 model in R and calculate the pairwise distance of our sequences in R. Recall that for $\hat{p}$ = *p-distance = Hamming distance / total length* the distance estimator for this model is $\hat{d} = -\frac{3}{4}log(1 - \frac{4}{3}\hat{p})$.

	orangutan	gorilla	human	chimp	mouse
orangutan	______				
gorilla	______	______			
human	______	______	______		
chimp	______	______	______	______	
mouse	______	______	______	______	______

c) Add an additional line to your code to check whether all your sequences are of the same length. If they are not, return a warning message. Test your function using the sequences *sequences_primates_unequal_length.fasta*.

d) Read the primates sequences provided into R and use the dist.dna function of the R-package APE to control the distance matrices calculated above (the p-distance model is called "raw" in this package) and additionally calculate the pairwise distance using the K80 model.

e) Compare the three different distance matrices that you obtained in a), b) and d).

Exercise 2: A distance based method to build a phylogenetic tree (UPGMA)

Objective: To understand one method of building a phylogenetic tree and implement it in R.

a) The following is a pseudocode for the construction of a distance based tree, given a distance matrix. Try to make sense of the pseudocode. You are not asked to compute nor implement it yet but understanding it, shows how phylogenetic trees are build with the UPGMA method.

Pseudocode:

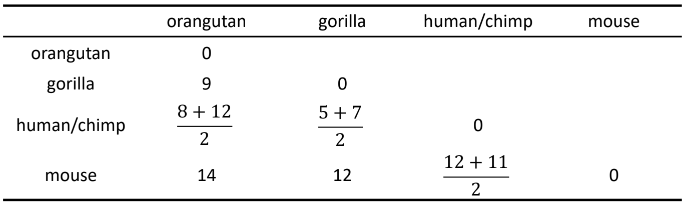
N_s : number of sequences in alignment (number of nodes);
N_c : number of clusters remaining to be merged;
$s = s_i$ for i in 1 to N_c : list of cluster descriptors/names;
$n = n_i$ for i in 1 to N_c :list of current size of each cluster;
branch$(x; y)$: branch length between nodes x and y;
leaf(x) : a tree leaf below the node x;
merge$(x; y)$: merging of two nodes in the tree;
D : Distance matrix;
$d[x; y]$: distance matrix entry for nodes x and y - distance between these nodes.

```
Data: Distance matrix D
Result: Ultrametric phylogenetic tree
for $i \leftarrow$ 1 to $N_s$ do
    $n_i \leftarrow 1$;
    $s_i \leftarrow node(i)$;
end
while $size(D) \geq (1; 1)$ do
    Choose $s_i, s_j$ such that $min(D) = d[s_i, sj]$;
    $n_{i,j} \leftarrow n_i + n_j$;
    $s_{i,j} \leftarrow merge(s_i, s_j)$;
    branch$(s_{i,j}, s_i) \leftarrow d[s_i, s_j]/2-$branch$(s_i, leaf(s_i))$;
    branch$(s_{i,j}, s_j) \leftarrow d[s_i, s_j]/2-$branch$(s_j, leaf(s_j))$;
    for $m \neq i$ and $m \neq j$ do
        $d[s_m, s_{i,j}] \leftarrow \frac{n_i d[s_i, s_m] + n_j d[s_j, s_m]}{n_i + n_j}$
    end
    delete $d[s_i, \cdot], d[\cdot, s_i]$; delete $d[s_j, \cdot], d[\cdot, s_j]$;
end
```

Hint 1: If this is difficult, check out an example on
http://www.evolution-textbook.org/content/free/tables/Ch_27/T9_EVOW_Ch27.pdf

b) Apply the algorithm to the distance matrix derived with Hamming distance in exercise 1 (do it manually).

Hint 2: In the first step you should cluster human and chimp and get a new distance matrix that looks like this



	orangutan	gorilla	human/chimp	mouse
orangutan	0			
gorilla	9	0		
human/chimp	$\frac{8+12}{2}$	$\frac{5+7}{2}$	0	
mouse	14	12	$\frac{12+11}{2}$	0

Hint 3: You can check your result using the upgma function of the phangorn package in R.

c) **Bonus:can be done after all exercises are completed** Implement the UPGMA algorithm in R (there is an R skeleton provided to you, where you “only” have to fill in the gaps) and plot the resulting tree.

Exercise 3: Application of phylogenetics: the origin of HIV

Objective: To make a phylogeny, to infer from it the origin of HIV therefore to see how HIV-1 and HIV-2 relate to SIV sequences.

In this part, you will be able to understand a great molecular epidemiology discovery. The file *HIV_SIV_alignment.fasta* contains and alignment of different HIV sequences as well as SIV (Simian Immunodeficiency Virus) sequences from different monkeys species.

To plot such a tree:

- Import the fasta file like in the R preparation box at the beginning of the exercises
- Use the function *dist.dna* with the *JC69 model* to get the distance matrix for the alignment
- You could visualize the distance matrix using the function *View* from the *utils* package or the function *heatmap* from the *stats* package
- Use the function *nj* to wich infers a neighbour-joining phylogenetic tree from distance matrix
- Finally plot the phylogeny tree obtained

a) Was HIV zoonosis a unique event, or was it recurrent?

b) Do all HIV subtypes have the same origin?

c) Which species are most likely to have caused transmission to humans?

Exercise 4: Application of phylogenetics: Phylogenetics in court

Objective: Apply the knowledge gained in a real-life example and think about the ethics of applying phylogenetics in such a case.

Consider the following case (adapted from E-mail exchange Tanja Stadler/David Hillis and Nadine Bachmann):
In 1994 the American Physician Dr. Schmidt repeatedly threatened to kill his girlfriend J. Trahan or “make it so that no man would want to be with her” if she left him, sometimes in front of other witnesses. The physician forcefully gave her what he claimed was a “vitamin B-12 injection for depression”, against her will. The victim experienced a bad reaction from the injection, but broke all contacts with the physician at that time. The following year, she married another individual, got pregnant, and went in for routine pre-natal testing. At that point, she found out that she was HIV-positive. She had no other risk factors, and neither the physician nor her new husband was HIV positive. The only thing that she could think of was the suspicious injection. She went to the police, who investigated and found the suspicious blood draw records (initially missing but then) found hidden in a closet of the physician’s office, under a stack of boxes. The blood draw records showed that the physician took blood from his HIV positive patient (and also a Hepatitis C-positive patient) on the day that he injected the victim. The victim was infected with both HIV and Hepatitis C. The blood draws were not sent to any lab for testing; nor were the patients charged for the blood draws. The HIV-positive patient was a homosexual man in a long-term monogamous relationship, and did not know the victim.

a) If you were the investigator in charge of this criminal case – how would you use phylogenetic tools as evidence in this case? Describe and discuss together.

b) Using the sequences provided and the phylogeny building procedure from the last exercises, what can you conclude?

c) Compare to how the analysis was actually done: https://doi.org/10.1073/pnas.222522599, *Metzker2002*

d) Discuss pros/cons of using phylogenetic tools in court. What prerequisites need to be given? Do you find the conviction of attempted second-degree murder adequate in the case discussed?

If there is still time you may either go back to the programming exercise 2.c) or continue with reading on more HIV criminal cases.

e) Discuss how phylogenetic tools were employed in the following two cases:

- Innocent nurses in Libya:
  https://en.wikipedia.org/wiki/HIV_trial_in_Libya#Scientific_studies_and_reports
  https://doi.org/10.1038/444836a, *deOliveira2006*
- Florida dentist:
  http://www.nytimes.com/1993/06/06/weekinreview/aids-and-a-dentist-s-secrets.html?pagewanted=all,
  AIDS and a Dentist's Secrets - The New York Times
  http://www.nature.com/nature/journal/v369/n6475/pdf/369024a0.pdf, *Hillis1994*

CIRCUMSTANTIAL EVIDENCE, THAT'S ALL YOU'VE GOT!
YOU CAN'T GET A FAIR TRIAL IN THIS TOWN.
© 1994 Watterson/Distributed by Universal Uclick
WATTERSON 1-23

Figure 1: Calvin and Hobbes: circumstantial evidence