Ensimag, Master MSIAM, Bio-Health Computing **2017-18**

Computational Biology

Phylogenetic analysis

Rules

Students should work in groups of 1 or 2 persons.

Your report should be sent by email to michael.blum@univ-grenoble-alpes.fr.

The deadline for sending your report is October 20, 2017. Penalties are incurred if not respecting the deadline.

Your report should be written using R Markdown `http://rmarkdown.rstudio.com/`, which is a file format to weave together narrative text and code (`R` scripts). R Markdown documents contain fully reproducible analyses that can be turned into high quality reports or presentations. The software RStudio provides a comprehensive IDE for `R` where you can write R Markdown documents (*.Rmd*) and convert them into *.pdf* documents. An R Markdown quick reference guide is available online. Reports should be sent both in *.Rmd* and *.pdf* formats. The report in *pdf* should contain 15 pages or less. If you do not know `R` and RStudio, you should read a brief introduction before the 2^{nd} session, which will take place on October, 10.

Introduction

In biology, phylogenetics is the study of evolutionary relationships among groups of organisms (e.g. species, populations), which are discovered through molecular sequencing data and morphological data (Wikipedia). A phylogenetic tree or phylogeny is a tree showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical or genetic characteristics (Figure 1). The taxa joined together in the tree are implied to have descended from a common ancestor. Phylogenetic trees are used for different purposes including classification of species and inference of the spatiotemporal evolution of epidemics. Well-explained examples of how evolutionary trees are used are provided at the Understanding Evolution website of the university of Berkeley (http://evolution.berkeley.edu/evolibrary/article/phylogenetics_01).

The objective of the 9 hour session is to show how to construct phylogenetic trees based on an alignment of DNA sequences. Here, we assume that the DNA molecule is composed of succession of nucleotides of four different types (A, C, G or T). For instance, for DNA sequences— obtained with multiple alignment—of 3 unknown species that would look like

```
Species 1 ACCCT
Species 2 ACGCT
Species 3 ACCGG
```

we can anticipate that species 1 and 2 are the most related because they differ at one nucleotide by contrast with species 3 that have more than one difference when compared to species 1 or 2. The differences between sequences arise because of *substitutions*,

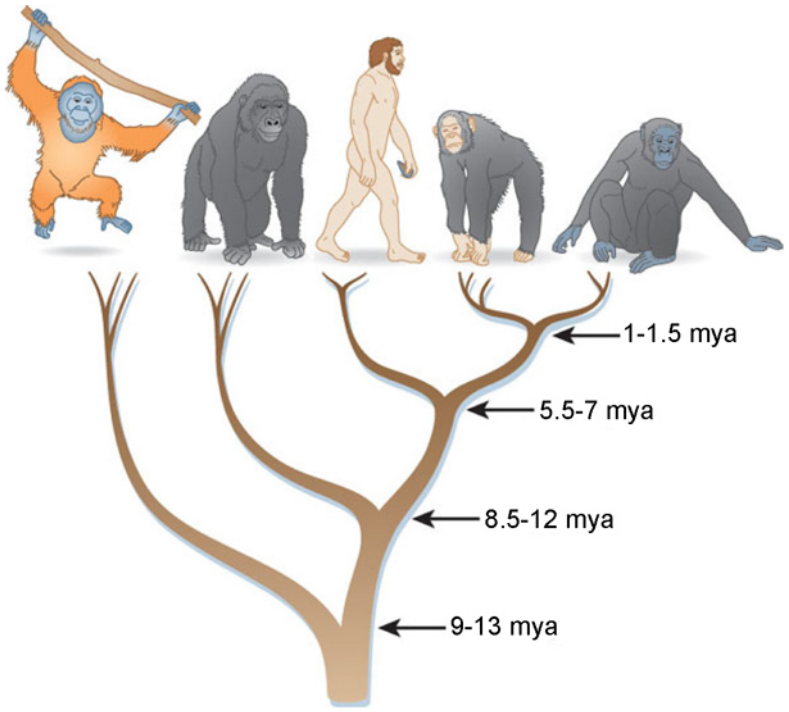

Figure 1: Evolutionary relationships of Hominidae excerpted from Mitchell and Gonder (2013). Chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) are the species most related with humans (*Homo sapiens*) followed by gorillas (*Gorilla gorilla*) and orang-outans (*Pongo abelii*). In the course, we will study how to reconstruct phylogenetic trees from DNA sequences.

which are mutations where one nucleotide is replaced by a different one (e.g. $A \to G$ in the 3^{rd} site of the just mentioned DNA sequences).

1. Draw a phylogenetic tree with three substitutions that is compatible with the 3 DNA sequences provided as an example.

In the following, we introduce distance-based methods in phylogeny. Distance-based methods consist of two steps which include 1. the computation of a distance matrix between species and 2. the construction of a phylogenetic tree based on a distance matrix.

Part 1. Computation of evolutionary distances

The simplest distance between two DNA sequences is the proportion of sites that differ between two sequences. This distance is useful when the number of substitutions per site is so small that there are virtually no multiple hits at each site. When multiple hits cannot be neglected, more refined versions of distances should be computed.

To compute new distances, we assume the Jukes-Cantor model of molecular evolution. It assumes that substitutions occur according to a Poisson process of rate ν (time is counted in years for instance) and that each of the three possible substitution is equally likely. The different sites are assumed to evolve independently from one another.

1. Provide a definition of a Poisson process and give at least 1 example (not necessarily in biology).

2. In an interval of length Δt, provide the probability that there is exactly one substitution and give a first-order approximation. Give also a first-order approximation of the probability that there is no mutation in an interval of length Δt.

3. We assume the Jukes-Cantor model for the evolution of a nucleotide. We denote by $f(t)$ the probability that the nucleotide is the same after a duration t (counted in years). There are different options for a nucleotide to be the same (e.g. an A remains an A) including no mutations during a duration t, or a first mutation at time s_1 (at $s_1 < t$, $A \to T$) followed by a back-mutation at time s_2 ($s_1 < s_2 < t$, $T \to A$).

   Compute $f(t + \Delta t)$ as a function of $f(t)$. To make the computations, you should consider two options depending or not the two nucleotides are the same at time t.

4. Based on question 3, show that f is the solution of an ordinary differential equation (ODE) and give the initial condition.

5. Solve the ODE. Show that the probability that the two nucleotide sites differ after a time t is given by

$$1 - f(t) = \frac{3}{4}(1 - e^{-4\nu t/3}). \tag{1}$$

6. (in front of a computer) For $\nu = 1$, plot the probability that two nucleotides differ as a function of time t. Explain why it does not increase linearly with time.

7. We change the time unit so that one unit of time corresponds to ν years ($t' = \nu t$). We assume that we have computed the proportion p of nucleotides that differ between two sequences. In the new time unit, propose an estimation of the evolutionary time that separates two DNA sequences. The obtained formula corresponds to the Jukes-Cantor distance. It is a widely used distance in phylogenetic analyses.

8. (in front of a computer) Plot the Jukes-Cantor distance as a function of the proportion of difference p between two sequences. For which values of p, does it matter to use the Jukes-Cantor distance instead of the naive proportion distance?

Part 2. Construction of phylogenetic trees

The UPGMA algorithm

The first method we introduce for constructing phylogenetic trees is the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). It is a clustering method where at each stage two clusters are merged to form a new cluster that corresponds to a node of the tree. If we denote by s_1, s_2, ..., s_n the species of the tree, it works as follows

- *Initialization*: Define n clusters C_1, ..., C_n each of them containing one species. Compute the distance between clusters as $d(C_i, C_j) = d(s_i, s_j)$.
- *Iteration* Merge the two clusters C_i and C_j that are the nearest according to the distance d. Form a new cluster C_m and compute the distance between cluster C_m and the other ones as follows

$$d(C_m, C_l) = \frac{d(C_i, C_l)\#C_i + d(C_j, C_l)\#C_j}{\#C_i + \#C_j},$$

  where $\#C_i$ is the number of species in cluster C_i. Cluster C_m corresponds to a new node in the tree that is ancestor of C_i and C_j, and the distance between C_m and its two descendants is equal to $d(C_i, C_j)/2$ when $\#C_i = \#C_j = 1$.

The main result about the UPGMA algorithm is that it is able to reconstruct the true tree when the distance between species is *ultrametric*. To define ultrametric distances, we first define *additive* distances. Additive means that there exists a tree such that the distance between any two species is the sum of the lengths of the branches connecting them. Ultrametric distances are additive distances such that the distance from the root of the tree to all leaves is the same. It means that the amount of evolutionary divergence from the common ancestor should be the same for all species, which is referred as the molecular clock hypothesis. A necessary and sufficient condition for a distance to be ultrametric (admitted) is the three point condition, which states that for any triplet of species s_i, s_j, s_k two distances between them are equal $d(s_i, s_j) = d(s_i, s_k)$ and the third is less than these two $d(s_j, s_k) \leq d(s_i, s_j)$.

1. Compute the algorithmic complexity of an implementation of the UPGMA algorithm.
2. Show that UPGMA corresponds to hierarchical clustering with a particular linkage function. Provide the corresponding `R` function.
3. For the tree displayed on the left panel of Figure 2, compute the distance between species that derived from the tree and run the UPGMA algorithm by hand.
4. Answer to the same question for the tree displayed in the right panel of Figure 2.
5. Provide an explanation for why the performance of UPGMA differs for the 2 trees.

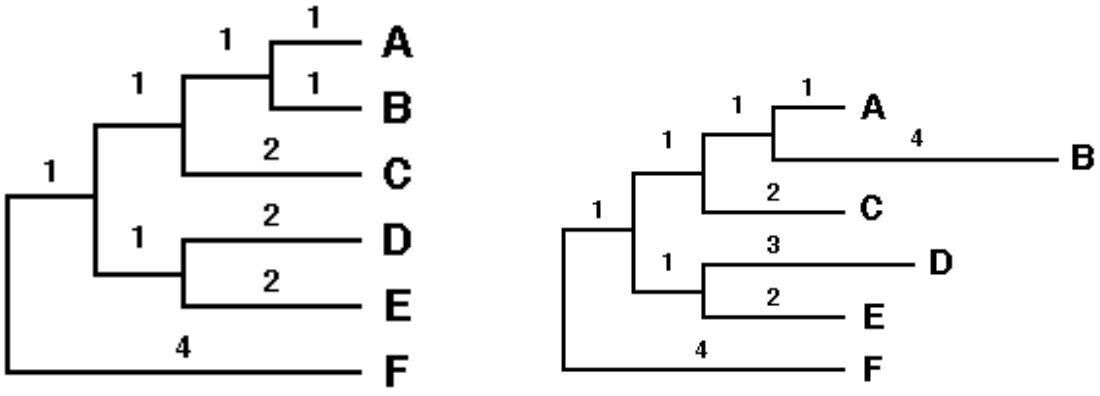

Figure 2: Two examples of phylogenetic trees.

The NJ algorithm

The second method we introduce for constructing phylogenetic trees is the Neighbor-joining (NJ) method (Saitou and Nei, 1987). It does not require that the distance is ultrametric. In other words it does not require that all lineages have diverged at the same rate from the common ancestor. The NJ method generates an unrooted tree and we now describe how it proceeds. The raw data are provided as a distance matrix and the initial tree is a star tree. Then a modified distance matrix is constructed in which the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes. The tree is constructed by linking the least-distant pair of nodes in this modified matrix. When two nodes are linked, their common ancestral node is added to the tree and the terminal nodes with their respective branches are removed from the tree. This pruning process converts the newly added common ancestor into a terminal node on a tree of reduced size. At each stage in the process two terminal nodes are replaced by one new node. The process is complete when two nodes remain, separated by a single branch.

The modified matrix is computed as follows

$$Q(s_i, s_j) = (n-2)d(s_i, s_j) - \sum_{k=1}^{n} d(s_i, s_k) - \sum_{k=1}^{n} d(s_j, s_k) \tag{2}$$

If species s_i and s_j are joined to form a new node s', the distance between s' and the species s_i and s_j which is reported in the constructed tee is obtained as

$$\delta(s_i, s') = \frac{1}{2}d(s_i, s_j) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(s_i, s_k) - \sum_{k=1}^{n} d(s_j, s_k)\right]$$

and

$$\delta(s_j, s') = d(s_i, s_j) - \delta(s_i, s').$$

The distances between the new node s' and the remaining species are computed as follows

$$d(s', s_k) = \frac{1}{2}[d(s_i, s_k) + d(s_j, s_k) - d(s_i, s_j)] \tag{3}$$

The main result about the NJ algorithm is that it is able to reconstruct the true tree when the distance between species is *additive*.

1. Compute the algorithmic complexity of the NJ algorithm.
2. Provide a (possibly graphical) explanation for why using equation (3).
3. Explain why using Q instead of d for merging species might account for lineages with varying rates of evolution (see right panel of Figure 2)
4. For the tree displayed on the right panel of Figure 2, run the NJ algorithm by hand and comment the result.

Part 3. Construction of phylogenetic trees with R

Install and load the R package *ape*

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

Load the sequence data

```
dat<-url("http://membres-timc.imag.fr/Michael.Blum/fasta.Rdata")
load(data)
```

Tree reconstruction

1. Does *data* contains sequence data for all the species of Figure 1?
2. Using the function *dist.dna*, compute the simple distance, which is the proportion of differences between sequences and the Jukes-Cantor distances. Compare the two distances and explain why there are (no) differences. Help about the *dist.dna* function can be obtained as follows

   ```
   ?dist.dna
   ```

3. Perform the UPGMA algorithm and compare the obtained phylogeny to the one of Figure 1.
4. Consider different linkage criteria to evaluate if it impacts the obtained phylogeny.
5. Perform the NJ algorithm and propose a root for the phylogeny. Compare to the phylogeny of Figure 1 and comment the result.
6. Would you recommend to use UPGMA or NJ in practice?

Bootstrap

1. Explain what is the bootstrap procedure in phylogeny.
2. Perform bootstrap using *boot.phylo* and superpose bootstrap values on the NJ tree using *nodelabels*.
3. The objective of this question is to write your own bootstrap function. Bootstrap values measures the statistical support of each bipartition. Provide a definition of bipartition. Implement your bootstrap function using the functions `prop.part` and `prop.clades`.
4. Provide statistical or biological arguments to evaluate if the bootstrap procedure is adequate to measure the support of the proposed phylogeny.

Implementation of NJ (difficult)

1. Implement your own R function to perform NJ.
2. Compare your R function to the *nj* function of the package *ape*.