# Algorithms in Computational Biology, 2024
# Exercise 3 - Continuous-time Markov model

Due date: 17/03/2024

Submission should include a tar file named $ex3.tar$, containing the following files:

- conservation_coef.py
- Other python files are optional.
- example.pdf

## Conserved sequences

Conserved sequences are regions in the genome that show low mutation rate. Many such sequences have an important role in the regulation of gene expression. TATA-box for example, is an important DNA motif, that is associated with binding of TBP (TATA-binding protein), a key subunit of the transcription machinery. You can imagine, that in case of mutation of this motif, it may lose its function, which in turn will reduce the expression level of that gene, which may have a lethal effect. Thus, during evolution, we will see sequences like this having a lower mutation rate compared to the "less important" genomic regions. In this exercise, you will be given a file that contains multiple sequences. The format of the file is "PHYLIP", this format is commonly used in studies of phylogenetic evolutionary trees. The first line contains the number of sequences and the length of the aligned sequences, separated by a tab. The next lines contain the name of each sequence, then tab, and then the aligned sequence itself. The aligned sequences may contain gaps. For example:

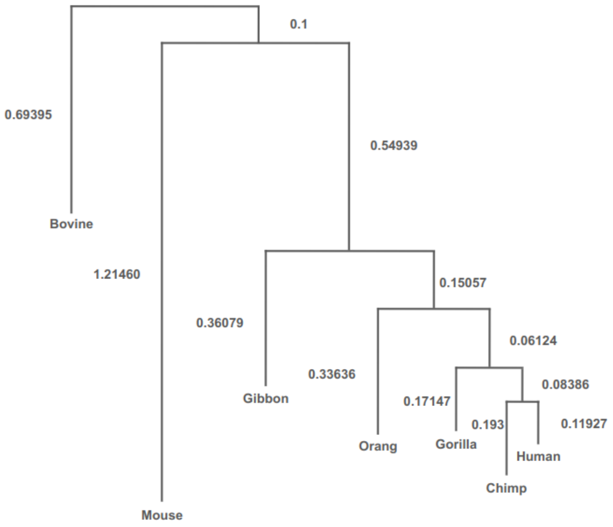
```
3       40
seq1    ACTCTGTCATCTTCTGAGATGTGACGCAGCGGATAACTAC
seq2    AGAGTGTCAAGAGAGTACAGGTGACGAACCCGATAACAAC
seq3    TCGTTGTCACGTGTCGCGTGGTGATGGATCGGATGACAAC
```

Together with the sequences, you will be provided with a file that represents a phylogenetic tree, in Newick format (see below). Using these two inputs, you are asked to determine how conserved each one of the positions in the alignment are. You may assume that every pair of nucleotides in the sequence is independent, meaning that mutation in one doesn't change the probability of mutation in another. In addition, assume that the tree is additive, meaning that the evolutionary distance between the nodes is the sum of edges that connect them, and it represents the similarity between the nodes. In the example below, the distance between a human and a chimpanzee is $0.193 + 0.11927 = 0.31227$ and the distance between a human and a gorilla is $0.11927 + 0.08386 + 0.17147 = 0.3746$. To determine the conservativity of a motif or a nucleotide, you have to find the conservation coefficient $\alpha$, that maximizes the likelihood of the input sequences given the phylogenetic tree, "inflated" or "shrunk" by some factor. The conservation coefficient is used to "inflate"/"shrink" all the edges of the tree. For example, $\alpha = 0$ represents maximal conservation, because it enforces the length of all edges of the deflated tree to be 0, meaning no mutation occurred between the sequences. You are expected to use a sliding window of length 11, and for each window to find the conservation coefficient of the subsequence in the window. Just to remind you, given a tree the calculation of the likelihood should be done for every position in the window and for every two nodes with an edge between them.

## Newick format

Newik format is a way to represent trees. Nodes are separated by commas, represented by strings or parentheses, followed by a colon and the length of the edge to the previous node. For example, the sequence

(Bovine:0.69395,(Gibbon:0.36079,(Orang:0.33636,(Gorilla:0.17147,(Chimp:0.193, Human:0.11927):0.08386):0.06124):0.15057):0.54939,Mouse:1.21460):0.10; represents the next phylogenetic tree:



## Visualisation of the inflation effect

For a window of your choise plot a graph that shows how the likelihood changes as a function of $\alpha$. In addition, explain in a few words what is shown on the graph and what was your way to find the best value for $\alpha$. Submit the graph and the explanation in a file named example.pdf.

## Specifications

You are allowed you use non-standard libraries (numpy, pandas, Bio) for reading/writing. Specifically, you may use the package Bio.Phylo for Python, together with io.StringIO if needed. It provides a convenient way to read trees in Newick format. In addition, you may use *scipy.linalg.expm* function. For the conservation coefficient assume that its range is $0 \leq \alpha \leq 0.1$. For the rate matrix please use the standard **Jukes-Cantor** matrix:

$$\begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}$$

It should be possible to invoke the program from the command line using the following format:

```
python3 conservation_coef.py tree.tree sequences.phylip
```

### Input

- *tree.tree* - File name that contains a string in Newick format, that represents the phylogenetic tree.
- *sequences.phylip* - File name of the aligned sequences in PHYLIP format.

### Output

The program should output the conservation coefficient for each one of the windows, meaning that the length of the output will be slighty shorter than the length of the sequences. For window of length 11, the output length will be 10 values shorter:

For example, if the input contains these sequences:

```
4       45
orig    AAAAAGCGCGGGCTTAAAAAAAAAATTTTATATAAAAAAAAAAAA
seq1    AAAAAGCACGGGCTTAAAAAAAAAACCGGCGCGGAAAAAAAAAAC
seq2    AAAAAAGGGGGGGGGAAAAAAAAAAGCGCGCGCCAAAAAAAAAAG
seq3    AAAAAGCGCGGGGTTAAAAAAAAAATATATATATAAAAAAAAAAT
```

And this tree:

```
((seq1:0.6, seq2:1.0)seq3:1.8)orig
```

Your output should look like (one column,3 digits precision, values separated by newlines):

```
0.039
0.039
0.062
0.075
0.088
0.1
0.088
0.075
0.062
0.05
0.05
0.05
0.05
0.029
0.019
0.029
0.05
0.074
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.062
0.029
0.029
```