ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Statistical Models in Computational Biology

Jack Kuipers
David Dreifuss
Xiang Ge Luo
Rudolf Schill

Due date: 30th March 2023
Please submit your project with the filename `Lastname(s)_Project5.pdf`.

Problem 12: Transition matrix, rate matrix, and stationary distribution **(1 point)**

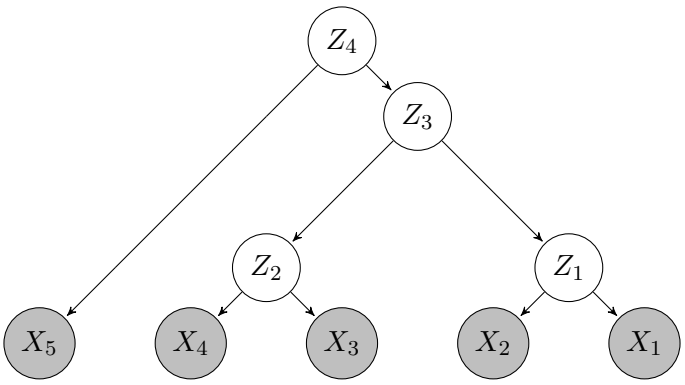
We can model a nucleotide substitution process by a continuous-time homogeneous Markov chain (see lecture notes). Let $\mathbf{P}(t)$ denote the transition matrix

$$\Big(P\big(y(t)=a \mid y(0)=b\big)\Big)_{a,b\in\{A,C,G,T\}}$$

where $\mathbf{P}(0)=\mathbf{I}$ (the identity matrix). For an infinitesimally small time interval $\mathrm{d}t$, we assume that $\mathbf{P}(\mathrm{d}t)=\mathbf{P}(0)+\mathbf{R}\mathrm{d}t$, where $\mathbf{R}$ is the *rate matrix*.

1. Show[1] that $\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t}=\mathbf{R}\mathbf{P}(t)$. (0.5 points)
2. Assume that the given Markov chain is ergodic with (unique) stationary distribution $\vec{\pi}$. Show that $\mathbf{R}\vec{\pi}=\vec{0}$. (0.5 points)

Problem 13: Phylogenetic trees as Bayesian networks **(3 points)**



Consider the above phylogenetic tree T, where the leaves $X=(X_1,\ldots,X_5)$ are the random variables that model the appearance of the nucleotides in an alignment of five sequences. Each of the internal nodes $Z=(Z_1,\ldots,Z_4)$ models the appearance of nucleotides of an ancestral sequence. We interpret the phylogenetic tree as a Bayesian network.

1. What is the joint probability $P(X,Z|T)$ of the tree? (1 point)

[1] *Hint:* Use the derivative's definition, $\mathbf{F}'(t):=\frac{\mathbf{F}(t+\mathrm{d}t)-\mathbf{F}(t)}{\mathrm{d}t}$, as well as Chapman–Kolmogorov's equation $\mathbf{P}(t+s)=\mathbf{P}(t)\mathbf{P}(s)$ for $t,s>0$.

2. How many summation steps would be required for the naive calculation of $P(X|T)$ via brute-force marginalization over the hidden nodes Z? (1 point)

3. Rearrange the expression $P(X|T)$ such that the number of operations is minimized [2]. How many summation steps are required now for the calculation of $P(X|T)$? (1 point)

Problem 14(data analysis): Learning phylogenetic trees from sequence alignment data (6 points)

The aim of this problem is to solve a forensic case by studying phylogenetic trees. For this purpose you will estimate the topology and the parameters of a phylogenetic tree from sequence alignment data using the R packages `phangorn` and `ape`. Learning a phylogenetic tree with these packages comprises the following steps:

i. Calculate pairwise distances between the sequences in the alignment.

ii. Generate a first tree topology based on neighbour joining.

iii. Optimise the likelihood of a phylogenetic model with respect to branch lengths, nucleotide substitution rates, and tree topology. Phangorn uses heuristics to locally search the space of possible tree topologies and to find a topology with (locally) highest score.

We will look at a dataset collected in a health care institution in the suburbs of Paris. Each data point is the DNA sequence of a certain genomic region ('RT') of the HIV-1 virus extracted from HIV-1 positive persons. This dataset was collected in the framework of a virological investigation[3] requested by the authorities in order to determine whether a nurse-to-patient infection had occurred in the health care institution mentioned above. A patient was infected with HIV-1 during her stay at the institution. After screening the members of the staff who had had contact to the patient, the authorities found out that one male nurse and one female nurse were already HIV-1 positive during the time the patient was at the institution. Your task is to establish which nurse is more likely to have infected the patient. Furthermore, you have to assess the likelihood that the patient was actually infected at the hospital based on the phylogenetic comparison to control HIV-1 strains from the Paris area. In order to assess the uncertainty associated with phylogenetic tree inference, you will compare the outcomes of the phylogenetic tree learning procedure across bootstrap resamples from the alignment.

1. Install and load the R packages `phangorn` and `ape`. Load the alignment `ParisRT.txt` into memory using the function `read.dna()`.

2. Create an initial tree topology for the alignment, using neighbour joining with the function `NJ()`. Base this on pairwise distances between sequences under the Kimura (1980) nucleotide substitution model, computed using the function `dist.dna()`. Plot the initial tree. (1 point)

3. Use the function `pml()` to fit the Kimura model (`model = "K80"`) to the above tree and the alignment. Note that the function expects `data = phyDat(alignment)`. What is the log likelihood of the fitted model? (1 point)

4. The function `optim.pml()` can be used to optimise parameters of a phylogenetic model. Find the optimal parameters of the Kimura (1980) nucleotide substitution model whilst the other parameters are held *fixed*. What are the values in the optimised rate matrix? (1 point)

[2] *Hint*: Find inspiration in the computation of $P(X)$ in hidden Markov models.

[3] Goujon C. et al. *Phylogenetic analyses indicate an atypical nurse-to-patient transmission of Human Immunodeficiency Virus Type 1*. Journal of virology, 2000

5. Optimise the Kimura model with respect to branch lengths, nucleotide substitution rates, and tree topology simultaneously. What is the log likelihood of the optimised model? (1 point)

6. The function `bootstrap.pml()` fits phylogenetic models to bootstrap resamples of the data. Run it on the optimised model from step 5, but pass the argument `optNni = TRUE` to allow for a different topology for each bootstrap run. What, exactly, is being resampled? (1 point)

7. Use `plotBS()` with `type = "phylogram"` to plot the optimised tree (from step 5) with the bootstrap support on the edges. Which nurse ("Mme_S" or "Mr_D") is more likely to have infected the patient "Mme_L"? (1 point)