Exercise 1: Calculate the equilibrium distributions of the Markov chains with transition matrices U and V and check whether the processes are reversible:

$$U = \begin{pmatrix} 0.3 & 0.7 \\ 0.1 & 0.9 \end{pmatrix} \qquad V = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.5 & 0.4 & 0.1 \\ 0.1 & 0.4 & 0.5 \end{pmatrix}$$

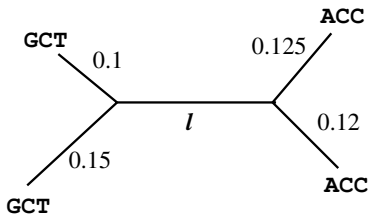
Exercise 2: Let $X_1, X_2, \ldots$ be a Markov chain on a finite state space $\mathcal{S}$ with transition probabilities $P_{x\to y}$ for all $x, y \in \mathcal{S}$. Let $(\pi_x)_{x\in\mathcal{S}}$ be a distribution that fullfills the detailed balance condition

$$\forall_{x,y\in S} : \pi_x \cdot P_{x\to y} = \pi_y \cdot P_{y\to x}.$$

(a) Show that $(\pi_x)_{x\in\mathcal{S}}$ is a stationary distribution of the Markov chain.

(b) Do stationary distributions of a Markov chain always fulfill the detailed balance condition? If yes, give a proof; if no, give a counter example.

Exercise 3: Find the central branch lenght ℓ that maximizes the likelihood of the tree shown below (for fixed lengths of the other branches). For the substitution process assume a Jukes-Cantor model with $\lambda = 1$, such that the rate of a change from nucleotide x to any *other* nucleotide y is $\lambda/4$. (Note: For this assignment you may want to write a e.g. an R or python script.)



Exercise 4: Assume an ancestral sequence and a derived sequence are given, and we aim to estimate the evolutionary distance between the sequences. For this, we assume a Jukes-Cantor model with rate $\lambda = 1$. Let n be the length of the sequences, and k be the number of segregating sites. (There are no gaps in the alignment or the alignment is known and positions with gaps are not counted.)

(a) Let $f(t)$ be the expectation value for the number of segregating sites if t is the true time distance between the sequences (assuming the Jukes Cantor model and sequence length n). Find a mathematical formula to express the function $f(t)$.

(b) Calculate the log-likelihood function $\ell_k(t)$, which is the log of the probability to observe k segregating sites if t is the right time (assuming Jukes Cantor etc.)

(c) The so-called *moment estimator* for t is the $\widetilde{t}$ such that $f(\widetilde{t}) = k$. How does it depend on the observed k? (It is called moment estimator because the expectation value is also called the first moment of a distribution.)

(d) As you know, the ML estimator $\hat{t}$ is the t that maximizes $\ell_k(t)$. How does $\hat{t}$ depend on k?

(e) Compare $\widetilde{t}$ to $\hat{t}$. Can you find an obvious relationship between the two? And if so, does it also hold for other substitution models that assume independence between the sites?

Exercise 5: Assume that 10,000 trees haven been MCMC-sampled from the posterior distribution for a given data set. To summarize these trees we would like to show a single (not necessarily binary) tree that has all branches that appear in more than 5,000 of the sampled trees (as splits of the taxa sets). Proof that this is always possible or present a counter-example.

Exercise 6: Simulate sequence datasets with different trees with 5 and with 10 taxa and with different mutation models (e.g. JC and HKY). Explore how the accuracy of the ML tree found by DNAML depends on the substitution model used in DNAML and the sequence length. Apply also RAxML to the data.

Exercise 7: Find a way to simulate data for a gene that has two exons and one intron. In the coding region, the third codon position evolves faster than the first two, and the intron evolves even faster. You can use seq-gen and postprocess the output. For simplicity, neglect that the mutation probabilities of the third codon position depend on the current states of the first two. Simulate several phylogenetic datasets and analyze them with RAxML with and without partitioning. How is the effect of partitioning on the accuracy of the result and how does this depend on the sequence lengths, the number of taxa, and the branch lengths of the tree?