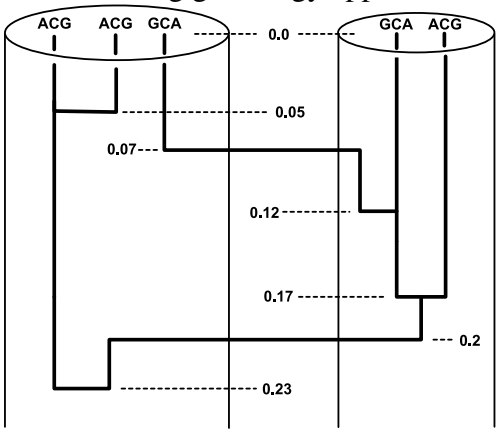
Dirk Metzler WS 2022/2023

COMPUTATIONAL POPULATION GENETICS — EXERCISE SHEET 3

1. The following genealogy appears during an MCMC sampling procedure in LAMARC.



The times are given in units of $1/\mu$ generations, where μ is the mutation rate per site. Population 1 (left) has an effective size of 10.000.000 individuals, and the effective size of population 2 (right) is 2.000.000 individuals. The sequences were sampled from a diploid locus. The mutation rate is $2.5 \cdot 10^{-8}$ per position and per generation. Per generation on average 0.8 individuals migrate from population 1 to population 2 and 0.1 individuals migrate from population 2 to population 1.

   (a) Calculate coalescence and migration rates (of ancestral lineages back in time) per time unit of $1/\mu$ generations.

   (b) Make a random proposal of a change in the tree as in LAMARC. That is, use a random generator (e.g. with the `sample` command in R) to randomly select one of the eight branches, delete it and use a random generator (e.g. the `sample` command and the `rexp` command in R) to let the ancestral lineage of the thus isolated sequence or subtree coalesce again with the rest of the genealogy.

   (c) Calculate the likelihood of the original genealogy and your modified genealogy for the sequence data given at the tips of the tree. Assume the Jukes-Cantor sequence evolution model.

   (d) Decide as in the Metropolis-Hastings algorithm in LAMARC whether your proposed modification is accepted or rejected.

   (e) Assume that the genealogy shown above and your modified tree were the only trees sampled from the posterior, where the parameter values as given above were assumed. Use importance sampling with these two trees to approximate the likelihood ratio of the given set of parameter values compared to the set of parameter values in which the effective size of population 2 is 4.000.000 individuals (and everything else as specified above).

2. In a cell line you study a certain locus that contains two CpG sites.
In each cell at a certain time point in cell cycle you distinguish three states: none of the two sites is methylated (0), one of the two is methylated (1), both are methylated (2). After many observations you infer that the transition probabilities $P_{x\to y}$ that a cell is in state y if its mother cell was in state x are as specified in the table.

$P_{x\to y}$		y		
		0	1	2
x	0	0.7	0.2	0.1
	1	0.1	0.7	0.2
	2	0.2	0.1	0.7

   (a) Calculate an equilibrium distribution for the Markov process of the number of methylated sites (where the cell analyzed in generation n is always an offspring of the cell analyzed in generation $n-1$).

   (b) Explain why or why not this Markov chain will converge against this equilibrium distribution.

(c) Analyze whether this equilibrium is reversible (that is, whether the detailed-balance condition is fulfilled).

(d) Simulate the process in R repeatedly, always starting in state 0, and analyze how probably state 0, 1 and 2 are for step n =1, 2, 3, ..., 20.

3. A certain random distribution on the positive integers $\{1, 2, 3, \ldots\}$ has the property that each number n is twice as probable as $n + 1$.

   (a) Specify a Metropolis-Hastings MCMC sampler to generate samples from this distribution. Assume that the only source of randomness that can be used is a function that gives you values from $\{-1, 0, 1\}$ for input probabilities q_{-1}, q_0, q_1 (where one of them can be 0).

   (b) Implement and test your algorithm in R, using as the source of randomness only commands of the form `sample(c(-1,0,1),1,p=q)` with suitable probability vectors `q`.

4. Use software like Tracer to evaluate the results of the preliminary LAMARC runs from exercises 6 and 7 of exercise sheet 2 and decide which MCMC options may be appropriate for the full analysis. Start LAMARC runs with these parameters.

5. Proof that any distribution p that fulfills the reversibility condition for an irreducible aperiodic Markov chain X on a finite state space must be a stationary distribution of X.

6. Simulate datasets for a population that consists of four sub-populations, with gene flow between the populations.

   (a) How accurately can you estimate the rates and favored directions of migration between each pair of sub-populations and how does this depend on the number of loci and other properties of the dataset?

   (b) Assume that each of the sub-populations has a substructure. How does this influence the results, especially if the the sampling was limited to some of the sub-subpopulations. What if the sub-substructure is ignored or unknown?

   Design the study and start the first preliminary test runs whose results we can discuss next week to decide about the program option settings for the rest of the study.

7. Explore with simulated datasets how one can increase the efficiency of MCMC Methods like LAMARC and IM/IMa by using MCMCMC and fine-tuning the heating parameters.