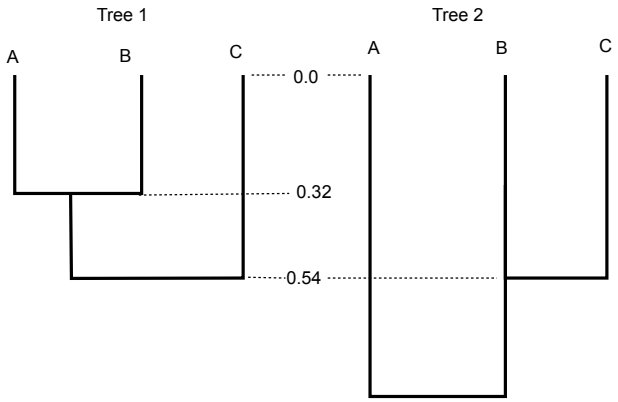
1. In the genealogy of the three sampled sequences named A, B and C, two recombination events took place after sequence positions 1234 and 2345. Tree 1 is the tree for sequence positions 1 to 1234 and tree 2 is tree for sequence positions 1235 to 2345.



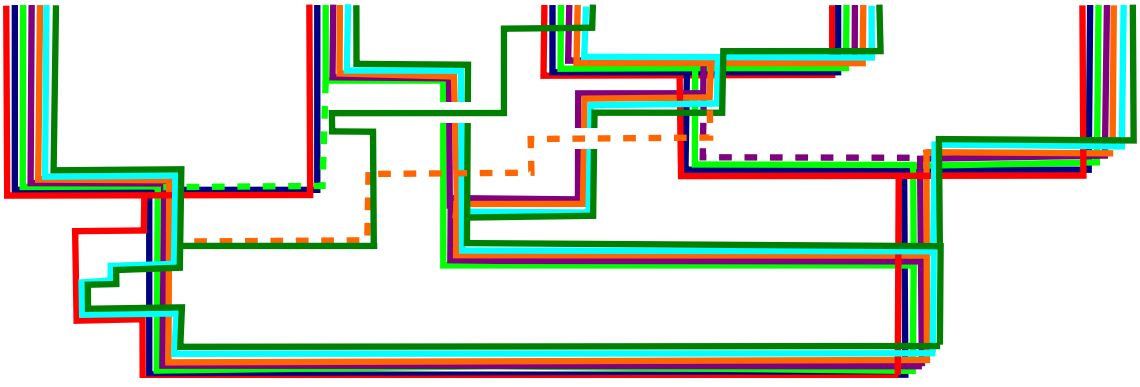
   Tree 3 is the tree from position 2346 on. Specify a set of trees such that the probability that tree 3 is among them has positive probability according to the ancestral recombination graph model but probability 0 according to the SMC' model.

2. Two sequences have been sampled from a diploid population. At sequence position $k-1$ the coalescence time is 1.2 units of $2N_e$ generations. The recombination rate between positions $k-1$ and k is $\rho = 4N_e r = 0.02$.

   (a) Given this information and assuming a constant population size, calculate for the SMC model the probability densities for the coalescence times 0.8 and 1.4 at position k.

   (b) Calculate, again for the SMC model, the corresponding probability densities for the case that the population has been exponentially growing in the last 1.1 time units, starting at a size of $\frac{1}{2}N_e$, which was the constant size before the exponential growth phase. (The time scaling is still in units of $2N_e$, where N_e is the present population size; Hint: The R function `integrate` might be useful.)

3. The recombination events in the following ancestral recombination graph (ARG) have taken place at sequence postiones 13423, 23402, 28873, 42031, 72345 and 96322. Dashed lines refer to non-ancestral material. (Note that not all non-ancestral lineages are shown and additional recombination events may have occurred on them.)



Which events in this ARG are neglected by...

(a) SMC ?

(b) SMC' ?

(c) MaCS (orignal version) with a threshold of 50000 bp?

(d) SCRM with a threshold of 50000 bp?

4. Loci A and B are on the same autosome with a recombination rate of $\rho = 4N_e r = 5$, where r is the recombination rate between the two loci per generation. Assume that the population is constant in size, panmictic etc. We trace back the lineages of the two loci from two copies of the chromosome (in a thought experiment as this may be difficult in practice), e.g. the two copies sampled from one individual. If nothing else is stated, assume for the following calculations the standard ancestral recombination graph (ARG) model.

   (a) Calculate the probability that the most recent event back in time was the coalescence of the two lineages and not a recombination event between loci A and B. (Here and in the following neglect within-locus recombination as well as mutation, indels etc.)

   (b) For the coalescent process of the lineages back in time, we distinguish the following states:

   u: There are two lineages and each of them is ancestral to a sample from locus A and a sample from locus B.

   v: There are three lineages. One is ancestral to a sample from locus A and a sample from locus B, one ancestral to a sample from locus A and the third is ancestral to a sample from locus B.

   w: All four lineages are separated.

   x: The lineages of locus A have coalesced and the lineages of locus B have not coalesced.

   y: The lineages of locus B have coalesced and the lineages of locus A have not coalesced

   z: The lineages of locus A have coalesced and the lineages of B have coalesced

   Which direct transitions between these states are possible and what are their probabilities?

   (c) Calculate the probability that the coalescent times at the two loci are exactly the same.

   (d) Calculate the probability that the coalescent times at the two loci are exactly the same, but this time for $\rho = 100$.

   (e) Calculate the probability that the two loci coalesce at the exact same time again for $\rho = 5$ and for $\rho = 100$, but this time with the SMC model instead of the ARG.

5. Check the correctnes of the version of the forward algorithm recursion with matrix–vector operations as given in the lecture and derive a similar version for the backward algorithm.

6. You have a shaker whith 3 dice that look exactly the same, but one of the tree dice is loaded and gives a six with probabolity 0.5 and each other result with probability 0.1. The other two dice are fair. You produce series of dice results in the following way: You pick random dice and roll it. Then you toss a fair coin and if it gives "head", you use the dice again. If the coin gives "tail", you put the dice back into the shaker, shake the three dice, pick one of them randomly and continue with that dice. This is repeated until you have a series of n results.

   (a) You do this with $n = 3$. Calculate the probability to obtain the sequence 2,3,6.

   (b) Specify a Hidden-Markov Model and all its parameters for this experiment.

   (c) With this procedure for $n = 15$ you obtain the series 1,3,2,5,6,6,2,6,1,3,4,5,1,6,3. Calculate the conditional probability that the 6 at the fifth position in this series came from the loaded dice, given the whole series.

   (d) Find the series of results of dice rolls in file cheater.txt. Calculate for each roll the probability that it came from the loaded dice, always conditioned on the whole series.

   (e) Conditioned on the whole series of results in cheater.txt, calculate the most probable series of "fair" and "loaded".

7. You sequenced a genomic region of 10000 bp. You expect that 10 % of the sites in this region belong to CpG islands of an average length of 100 bp. In CpG island, the probability that a C is followed by a G is 1/3, and for A, G or T the probability to be followed by a G is 2/9. Outside of CpG islands the probability that a C is followed by a G is 1/10, and the probability of A, G, or T to be followed by G is 0.3. In any case, the probability that A is the next nucleotide is always the same as the probability that it is C and as the probability that it is T.

   (a) Calculate the equilibrium distributions of nucleotides for CpG islands and for the other regions.

   (b) Specify a hidden Markov model to detect CpG islands in this genomic region and specify all parameters of the HMM.

   (c) The sequence of the genomic region is given in file cpg_islands.txt. Calculate for each position the probability of being in a CpG island (conditioned on the whole sequence).