

COMPUTATIONAL POPULATION GENETICS — EXERCISE SHEET 6

1. With the help of a dice and/or random generators available in R simulate a Moran model for a haploid population of only six individuals with selection.
2. With the help of a random generator in R simulate an Ancestral Selection Graph on a piece of paper, assuming a sample size of six with weak selection (and no recombination).
3. Simulate with *coala*/*msms* population genetic data containing selective sweeps and assess the sensitivity of Kim and Nielsen's ω and of *iHS*, *SDS* and *nS_L* to detect the sweeps. Also explore under what conditions demographic effects such as bottlenecks can lead to false positives in sweep detection.
4. You play a dice game against a cheater. You find out that the cheater began with a fair dice. But always when it was his turn to roll the dice and if he had used the fair dice before, he switched to a loaded dice with a probability of 0.05. He then used the loaded dice a geometrically distributed number of times (but at least once), and on average ten times. The loaded dice shows a six with probability 0.5 and each other result with probability 0.1.
 - (a) Specify a hidden Markov model and all its parameters for this strategy.
 - (b) Find the series of results of the cheater's 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.
5. 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 a 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).
6. Perform a simulation study to assess the accuracy and performance of PHASE. Start with a very simple model for simulating the input data. Then refine the model step by step.