D. Metzler, WS 2014/2015

COMPUTATIONAL POPULATION GENETICS — EXERCISE SHEET 5

1. Assume you simulate an ancestral recombination graph (ARG) for a n whole chromosomes sampled from a neutral, constant-size population (with positive recombination rate). From all the trees that this ARG assigns to single nucleotide positions, you choose four:

   Tree A is the tree at position 1000.

   Tree B is the next tree after Tree A, that is, you move on from position 1000 until there is a recombination event somewhere on a branch of tree A. Tree B is then the tree of the nucleotide position right after the recombination event.

   Tree C is the 100th tree in the list of all trees appearing from left two right, that is, the tree after the 100th recombination event (where only recombination events that effect the current tree are counted).

   Tree D is the next tree after Tree C, that is, you move on until there is a recombination event somewhere on a branch of tree C. Tree D is then the tree of the nucleotide position right after the recombination event.

   Which of the following statements are true? Substantiate your answers either logically or with computer simulations. (Start with $n = 2$ and $n = 3$). If you rely on computer simulations, try to find logical explanations for your observations.

   (a) The probability distribution of Tree A is that of a standard Kingman coalescent.

   (b) The probability distribution of Tree B is that of a standard Kingman coalescent.

   (c) The probability distribution of Tree C is that of a standard Kingman coalescent.

   (d) Tree B has the same probability distribution as tree A.

   (e) The expected total branch lenght of A is smaller than that of B.

   (f) The expected total branch lenght of B is smaller than that of A.

   (g) Tree C has the same probability distribution as tree A.

   (h) The expected total branch lenght of A is smaller than that of C.

   (i) Tree D has the same probability distribution as tree C.

   (j) The expected total branch lenght of C is smaller than that of D.

   (k) The expected total branch lenght of D is smaller than that of C.

2. Proof the following property of the Dirichlet distribution family:

   Let $N = (n_1, \ldots, n_K)$ be multinomially distributed with (unknown) probabilities $P = (p_1, \ldots, p_K)$, i.e.

$$\Pr(N = (n_1, \ldots, n_m)) = \frac{(n_1 + n_2 + \cdots + n_k)!}{n_1! \cdot n_2! \cdots n_k!} \prod_{i=1}^{k} p_i^{n_i}.$$

   If the prior distribution of P is $\mathcal{D}(\lambda_1, \ldots, \lambda_k)$, then the posterior distribution of P given $N = (n_1, \ldots, n_k)$ is

$$\mathcal{D}(\lambda_1 + n_1, \ldots, \lambda_k + n_k).$$

3. Consider a microsatellite evolution model with the following properties:

   - the equilibrium distribution of the repeat numbers is approximately normal with given mean μ and variance σ^2
   - Some mutations increase or decrease the number of repeats by one
   - Other mutations lead to a repeat number that is sampled from the equilibrium, independently of the previous state

   Develop a program that reads a tree in newick format, simulates the evolution of microsatellite repeat numbers along the tree and outputs the repeat numbers corresponding to the tips of the tree.

4. Simulate data and test the STRUCTURE software for several conditions (with and without admixture, with and without information about the sampling locations). In particular, consider two situations:

   (a) Three subpopulations have been seperated for many generations and recently started to exchange migrants.

   (b) Three subpopulations arose from a common ancestral population many generations ago, but there has always been some amount of gene flow between the populations. (Also try with more than three subpopulations.)

   (c) There are $N > 5$ subpopulations $1, 2, \ldots, N$, and there has always been geneflow. But direct geneflow between subpopulations i and j happened only if $|i - j| = 1$.

   Also try STRUCTURE runs assuming more subpopulations (K) than assumed when simulating the data.

5. 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.