D. Metzler, WS 2013/2014

COMPUTATIONAL POPULATION GENETICS — EXERCISE SHEET 5

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

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

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

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

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

2. Develop 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

   If you are a Bioinformatician, 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.

3. 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 $1, 2, \dots, N$, 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.