

If you have never used R or have no experience with computer programming, you may complete only problem 1 and 2 for full credit for the assignment. Others should complete all four problems.

1. The stationary distribution for the nucleotide bases, $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, is an array of four probabilities which sum to one. The Dirichlet distribution is a natural prior on this probability distribution. The Dirichlet density is

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \left(\frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha_i)} \right) \left(\prod_i \pi_i^{\alpha_i - 1} \right) \quad \text{with } 0 \leq \pi_i \leq 1 \text{ for all } i, \sum_i \pi_i = 1, \text{ and } \alpha = \sum_i \alpha_i$$

and where the density is determined by parameters $\boldsymbol{\alpha} = \{\alpha_i\}$ that must be positive. When $\alpha_i = 1$ for all i , this distribution is uniform over the space of possible distributions. For large α , the distribution is approximately multivariate normal.

The paper Larget and Simon (1999) describes an update method for $\boldsymbol{\pi}$ in which the proposal distribution is Dirichlet and has the current distribution as its mean. In other words, the proposal density q satisfies $q(\cdot \mid \boldsymbol{\pi}) = p(\cdot \mid \boldsymbol{\alpha} = A \times \boldsymbol{\pi})$ where A is a tuning parameter.

Assume a flat prior distribution for $\boldsymbol{\pi}$ (unnormalized density $h(\boldsymbol{\pi}) = 1$ for all valid $\boldsymbol{\pi}$). Using the Dirichlet update, find a mathematical expression for the acceptance probability when the current state is $\boldsymbol{\pi}$ and the proposed state is $\boldsymbol{\pi}^*$.

2. Write a program to implement the update method from the previous problem. Run your program to sample from the prior. Describe the behavior of your chain for different sets of pseudo-random numbers and different values of the tuning parameter A . What goes wrong?
3. Modify the proposal method from Problem 1 so that $q(\cdot \mid \boldsymbol{\pi}) = p(\cdot \mid \boldsymbol{\alpha} = (A \times \boldsymbol{\pi}) + \mathbf{1})$ where $\mathbf{1}$ is an array of ones. Find the new acceptance probability for this update. Implement this proposal mechanism. Does this modification fix the problem you observed in Problem 2?
4. The Tamura-Nei model is described on pages 200–203 in *Felsenstein*. Use the parameterization for the Q matrix where the expected number of nucleotide substitutions is one, so in addition to the stationary distribution $\boldsymbol{\pi}$, the model has free parameters R (the transition/tranversion ratio) and ρ (the ratio of purine transition rate over pyrimidine transition rate), where equations (13.7–13.9) show how to find values of parameters used in the Q matrix in Table 13.1 on page 201.

Use the data matrix below that summarizes the frequency of the 16 possible site patterns for the two-taxon Cow/Pig tree in the 1140 bp cytochrome b gene to find point estimates for $\boldsymbol{\pi}$, R , and ρ . (These need not be MLE estimates — simple estimates based on observed proportions suffice.) Implement an MCMC approach for sampling from the posterior distribution of the edge length of the two-taxon Cow/Pig tree under the Tamura-Nei model assuming an exponential prior on the edge length. Is the posterior sensitive to choice of prior?

For extra credit (whatever that means!) instead of using point estimates of the Q parameters, update them as well in your program.

		Cow			
		a	c	g	t
Pig	a	297	38	18	6
	c	32	252	2	42
	g	14	3	131	2
	t	13	51	2	237