Probability Model in BADGER

Prior for the Tree Topology.— The mathematical representation of an unrooted phylogeny for S taxa includes an unrooted tree topology τ. The unrooted tree topology is a connected acyclic graph with S labeled leaf nodes (each of which is adjacent to one other node in the tree), $I = S - 2$ unlabeled internal nodes (each of which is adjacent to three other nodes), for $N = I + S$ nodes in all, as well as a total of $E = 2S - 3$ edges. We let T_S represent the set of all such possible unrooted tree topologies with S leaves.

It is well-known that the number of such unrooted tree topologies is a product of increasing odd integers. We define $u(S)$ to be the number of unrooted tree topologies with S leaves, where

$$u(S) = \begin{cases} 1 & \text{for } S = 1, 2 \\ (2S-5)!! = 1 \times 3 \times \cdots \times (2S-5) & \text{for } S = 3, 4, \ldots \end{cases}$$

The number of rooted binary trees with S leaves, $r(S)$, is the number of unrooted trees with $S + 1$ leaves, namely $r(S) = u(S + 1)$.

In BADGER, the prior distribution is uniform on the set of unrooted trees that are consistent with a pre-specified partitioning of the taxa into groups, meaning that the tree has an edge for each group that separates the group from all other taxa. The default grouping is to place all taxa in a single group and to assume a uniform prior distribution on the complete set of $u(S)$ tree topologies. A partitioning of taxa into k groups of sizes $S_1, \ldots, S_k$ where each S_i is a positive integer and $\sum_{i=1}^{k} S_i = S$ results in a prior on a restricted set of tree topologies. Picking such a tree uniformly at random involves picking an unrooted tree topology for the number of groups and a rooted tree topology for each group. Specifically, the number of unrooted tree topologies subject to a partition is

$$u(k) \times \prod_{i=1}^{k} r(S_i) \ .$$

Prior Probabilities of Clades.— It is useful to point out that a uniform prior distribution on all tree topologies induces a prior probability distribution on unrooted clades (taxa separated from other taxa in the tree by a single edge) that is the same for all clades of a given size, but differs with clade size. In particular, the prior probability that a group of c taxa form a clade in an unrooted tree topology with S taxa is $r(c)r(S-c)/u(S)$. Prior probabilities for clades of size c and $S - c$ are equal and are larger when c is closer to 1 or S than when c is close to $S/2$.

Prior Distribution on Edge Lengths.— A vector of edge lengths $t = \{t_i\}$, for $i = 1, \ldots, 2S-3$ represents the expected number of gene inversions per edge. We assume that these means are independent and identically distributed Gamma random variables. The conventional parameterization of the Gamma distribution includes a shape parameter α and a scale parameter λ, both of which are positive, and has density

$$p(t \,|\, \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \mathrm{e}^{-\lambda t} \qquad \text{for } t > 0.$$

The re-parameterization $\mu = \alpha/\lambda$ and $\psi = 1+1/\lambda$ will be useful for expressing the prior distribution of the number of inversions per edge.

Prior Distribution on Inversions per Edge.— The conditional prior probability of x inversions on an edge given the length is Poisson(t), so that

$$p(x \mid t) = \frac{\mathrm{e}^{-t} t^x}{x!} \qquad \text{for } x = 0, 1, \ldots .$$

Integrating the Poisson probability versus the gamma prior results in the unconditional distribution

$$\mathrm{P}(x \mid \alpha, \lambda) = \frac{\Gamma(x+\alpha)}{\Gamma(\alpha) x!} \left(\frac{\lambda}{1+\lambda}\right)^{\alpha} \left(\frac{1}{1+\lambda}\right)^{x}, \qquad x = 0, 1, 2, \ldots .$$

Using the re-parameterization, the prior distribution for the number of inversions on an edge becomes

$$\mathrm{P}(x \mid \mu, \psi) = \left(\frac{\Gamma((\mu/(\psi-1)) + x)}{x! \Gamma(\mu/(\psi-1))}\right) \left(\frac{1}{\psi}\right)^{\mu/(\psi-1)} \left(\frac{\psi-1}{\psi}\right)^{x}, \qquad x = 0, 1, 2, \ldots .$$

This is a negative binomial distribution. The distribution has mean μ and variance $\mu\psi$. The parameter ψ measures the inflation of the variance of the negative binomial distribution relative to the the Poisson distribution with the same mean.

Prior Distribution on Reversals.— The biological process of gene inversion is represented by a reversal of a signed permutation where the reversal reverses both the sign and the order of a portion of the permutation. Each edge of the tree contains a list of reversals and their positions. Given the counts of reversals on the edges $x = \{x_i\}$, $i = 1, \ldots, E$, all reversals are mutually independent and selected uniformly from the set M_n of possible reversals that act on permutations of size n. There are $R(n) = \binom{n+1}{2} = n(n+1)/2$ possible reversals. We use the notation r_{ij} to denote the jth reversal on the ith edge which is located a distance u_{ij} from the beginning node. Distances are mutually independent chosen uniformly at random along the edge.

Summary of Prior Distribution.—

$$\begin{aligned}
\tau &\sim \text{Uniform}(T_S) \\
t_i \mid \mu, \psi &\sim \text{i.i.d. Gamma}(\alpha = \mu/(\psi-1), \lambda = 1/(\psi-1)) \quad \text{for } i = 1, \ldots, E \\
x_i \mid t_i &\sim \text{i.i.d. Poisson}(t_i) \quad \text{for } i = 1, \ldots, E \\
r_{ij} \mid x_i &\sim \text{i.i.d. Uniform}(M_n) \quad \text{for } i = 1, \ldots, E,\ j = 1, \ldots, x_i \\
u_{ij} \mid t_i, x_i &\sim \text{i.i.d. Uniform}(0, t_i) \quad \text{for } i = 1, \ldots, E,\ j = 1, \ldots, x_i
\end{aligned}$$

The joint prior on these parameters conditional on the hyper-parameters is

$$p(\tau, t, x, r, u \mid \mu, \psi) = \frac{1}{u(S)} \prod_{i=1}^{E} p(t_i \mid \mu, \psi) p(x_i \mid t_i) \left(\frac{1}{t_i}\right)^{x_i} \left(\frac{1}{R(n)}\right)^{x_i} .$$

A realization of all of these variables determines the observed data D, the set of observed genome arrangements, up to an arbitrary labeling.

Posterior Distribution of the Tree Topology.— We are primarily interested in evaluating the posterior distribution of the tree topology, $p(\tau \mid D, \mu, \psi)$, for given hyper-parameters. Bayes theorem tells us that the posterior is proportional to the likelihood and the prior.

$$\begin{aligned} p(\tau, t, x, r, u \mid D, \mu, \psi) &\propto p(\tau, t, x, r, u \mid \mu, \psi) p(D \mid \tau, t, x, r, u, \mu, \psi) && (1) \\ &= \frac{1}{u(S)} \prod_{i=1}^{E} p(t_i \mid \mu, \psi) p(x_i \mid t_i) \left(\frac{1}{t_i R(n)} \right)^{x_i} 1_{\{(\tau,x,r) \hookrightarrow D\}} && (2) \end{aligned}$$

where $1_{\{(\tau,x,r) \hookrightarrow D\}}$ is an indicator that D is consistent with the complete reversal history. We find the unnormalized posterior for τ, x, and r by integrating out the continuous parameters.

$$p(\tau, x, r \mid D, \mu, \psi) \propto \frac{1_{\{(\tau,x,r) \hookrightarrow D\}}}{u(S)} \left(\frac{\psi - 1}{\psi R(n)} \right)^{\sum_{i=1}^{E} x_i} \left(\frac{1}{\psi} \right)^{\mu E/(\psi - 1)} \prod_{i=1}^{E} \frac{\Gamma((\mu/(\psi - 1)) + x_i)}{x_i! \Gamma((\mu/(\psi - 1)))} \, .$$

We sample from this unnormalized posterior $p(\tau, x, r \mid D)$ using Markov chain Monte Carlo. The long-run relative proportion with which each tree topology appears converges to its posterior probability, $p(\tau \mid D, \mu, \psi)$.