5

Variable tree topology stepping-stone marginal likelihood estimation

Mark T. Holder
Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas, USA

Paul O. Lewis
Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA

David L. Swofford
Department of Biology and Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, USA

David Bryant
Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

CONTENTS

5.1 Introduction

The marginal likelihood is central to Bayesian model selection. It is the normalizing constant in Bayes' formula, and the Bayes factor used to compare two models is a ratio of marginal likelihoods. The marginal likelihood is defined as the expected value of the likelihood with respect to the prior. Methods for accurately estimating the marginal likelihood in phylogenetics were developed only recently (Lartillot and Philippe, 2006a; Xie et al., 2011; Fan et al., 2011; Arima and Tardella, 2012). Two of these methods—thermodynamic integration (TI; Lartillot and Philippe, 2006a) and the stepping-stone method (SS; Xie et al., 2011)—allow marginal likelihood estimation when the tree topology varies, but the most efficient methods to date—generalized stepping-stone (GSS; Fan et al., 2011) and the inflated density ratio method (IDR; Arima and Tardella, 2012)—have thus far remained restricted to estimating marginal likelihoods for a fixed tree topology. This chapter is concerned with updating GSS to allow variable tree topology, and the chapter by Wu et al. (Chapter 6) is concerned with updating the IDR method to allow variable tree topology.

5.2 The generalized stepping-stone (GSS) method

The goal of the GSS method (Fan et al., 2011) is to estimate the marginal likelihood,

$$p(\mathbf{y}) = \int_{\Omega} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where $\boldsymbol{\theta} = (\theta_g : g = 1, \cdots, n_g)$ is a vector of substitution model parameters, $\Omega \in \mathbb{R}^{n_g}$, and $\mathbf{y} = (y_j : j = 1, \cdots, n_j)$ is a vector of site patterns $y_j = (y_{jl} : l = 1, \cdots, n_l)$ where y_{jl} represents the single nucleotide state observed at site j for taxon l. The GSS method works by recognizing that estimating $p(\mathbf{y})$ is equivalent to estimating the ratio c_1/c_0, where c_β is the normalizing constant for a power posterior distribution of the form

$$p_\beta(\boldsymbol{\theta}) = \frac{q_\beta(\boldsymbol{\theta})}{c_\beta} = \frac{[p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})]^\beta p_0(\boldsymbol{\theta})^{1-\beta}}{c_\beta}.$$

Note that c_1 is the marginal likelihood of interest and c_0 is the normalizing constant for the arbitrary reference distribution $p_0(\boldsymbol{\theta})$. In the GSS method, $c_0 = 1$ because $p_0(\boldsymbol{\theta})$ is assumed to be proper.

The ratio $r = c_1/c_0$ is equivalent to following product of n_k ratios,

$$r = \frac{c_1}{c_0} = \left(\frac{c_{\beta_{n_k}}}{c_{\beta_{n_k-1}}}\right)\left(\frac{c_{\beta_{n_k-1}}}{c_{\beta_{n_k-2}}}\right)\cdots\left(\frac{c_{\beta_2}}{c_{\beta_1}}\right)\left(\frac{c_{\beta_1}}{c_{\beta_0}}\right),$$

where $\beta_k = k/n_k$, $k = 0, \cdots, n_k$. Each individual ratio $c_{\beta_k}/c_{\beta_{k-1}}$ composing this product can be estimated accurately using importance sampling, with $c_{\beta_{k-1}}$ serving as the importance distribution. Given MCMC samples $\{\boldsymbol{\theta}^{(i)}_{\beta_{k-1}} : i = 1, \cdots, n\}$ from $p_{\beta_{k-1}}(\boldsymbol{\theta})$, $r_k = c_{\beta_k}/c_{\beta_{k-1}}$ may be estimated as follows (Fan et al., 2011):

$$\hat{r}_k = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p\left(\mathbf{y}|\boldsymbol{\theta}^{(i)}_{\beta_{k-1}}\right)p\left(\boldsymbol{\theta}^{(i)}_{\beta_{k-1}}\right)}{p_0\left(\boldsymbol{\theta}^{(i)}_{\beta_{k-1}}\right)}\right]^{\beta_k-\beta_{k-1}}.$$

Numerical stability is improved by factoring out the largest term,

$$\eta_k = \max_{1\le i\le n}\left\{\frac{p\left(\mathbf{y}|\boldsymbol{\theta}^{(i)}_{\beta_{k-1}}\right)p\left(\boldsymbol{\theta}^{(i)}_{\beta_{k-1}}\right)}{p_0\left(\boldsymbol{\theta}^{(i)}_{\beta_{k-1}}\right)}\right\},$$

yielding the following estimator of the log marginal likelihood:

$$\begin{aligned}\log\hat{r} &= \sum_{k=1}^{n_k}\log\hat{r}_k \\ &= \sum_{k=1}^{n_k}\left[(\beta_k-\beta_{k-1})\log\eta_k\right] \\ &\quad + \sum_{k=1}^{n_k}\log\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{p\left(\mathbf{y}|\theta^{(i)}_{\beta_{k-1}}\right)p\left(\theta^{(i)}_{\beta_{k-1}}\right)}{\eta_k p_0\left(\theta^{(i)}_{\beta_{k-1}}\right)}\right)^{\beta_k-\beta_{k-1}}\right].\end{aligned}$$

In Fan et al. (2011), the topology was fixed and $p_0(\boldsymbol{\theta})$ had the same form as the joint prior distribution, with individual components adjusted so that their means and variances matched the corresponding sample means and variances from a pre-existing posterior sample. For example, two parameters may be estimated for the K80 model applied to just 2 sequences: the transition/transversion rate ratio, κ, and the edge length, ν (evolutionary distance between the 2 sequences). Assume that the joint prior density is a product of two Gamma densities (one for ν and the other for κ). If the marginal posterior distribution of ν has sample mean = 0.02 and sample variance = 0.0002, and the marginal posterior distribution of κ has sample mean = 4.0 and sample variance = 0.2, the reference distribution density would be a product of a Gamma(2.00, 0.01) density (for ν) and a Gamma(80.00, 0.05) density (for κ). Here, we generalize GSS further by incorporating the tree topology into the reference distribution.

5.3 Reference distribution for tree topology

In the previous discussion, $\boldsymbol{\theta}$ included all model parameters, including edge lengths. The fact that edge lengths are specific to a particular tree topology T requires modification of notation: $\boldsymbol{\nu}_T$ is now a vector of edge length parameters specific to tree topology T, and $\boldsymbol{\theta}$ now includes all other model parameters. The total parameter space $\Omega \in \mathbb{R}^{n_g + n_t n_b}$, where n_g is now the number of non-tree-specific model parameters and, for unrooted tree topologies with n_l tips,

$$\begin{aligned} n_t &= \frac{(2n_l-5)!}{2^{n_l-3}(n_l-3)!} \quad \text{(number of tree topologies)}, \\ n_b &= 2n_l - 3 \quad \text{(edge lengths/tree topology)}. \end{aligned}$$

The GSS method achieves its greater efficiency (over the SS method described by Xie et al., 2011) by using a reference distribution that is closer to the posterior distribution than the prior. As pointed out by Fan et al. (2011), maximum efficiency is obtained if the reference distribution equals the posterior exactly, in which case 1 MCMC sample is sufficient for estimating the marginal likelihood. While this maximum efficiency requires exact knowledge of the very quantity being estimated (the marginal likelihood) and is thus unobtainable, it is desirable to choose a reference distribution that is as similar to the posterior as possible. In addition to being close to the posterior, the reference distribution must be normalized because the GSS method assumes that $c_0 = 1$. Finally, the reference distribution should ideally allow direct sampling (rather than requiring Metropolis-Hastings updates).

In the next section, we describe a reference distribution that possesses all of these desirable properties, and we present an algorithm for sampling trees from this reference distribution. The proposed reference distribution, $p_0(T, \boldsymbol{\nu}_T, \boldsymbol{\theta}) = \pi(T) f(\boldsymbol{\nu}_T|T) f(\boldsymbol{\theta})$, is parameterized using a sample from a preliminary MCMC analysis, called the pilot run, exploring the posterior distribution. The goal is a distribution that samples trees roughly in proportion to the posterior distribution but does not rule out trees not visited in the pilot run. The general approach is:

1. draw a tree topology, T, from a distribution, $\pi(T)$, over all n_t bifurcating topologies (Section 5.3.1);
2. draw $\boldsymbol{\nu}_T$, a vector of n_b edge lengths, from $f(\boldsymbol{\nu}_T|T)$ (Section 5.3.2); and
3. draw substitution model parameters, $\boldsymbol{\theta}$, from $f(\boldsymbol{\theta})$ (Fan et al., 2011).

5.3.1 Tree topology reference distribution

Let a tree topology $T = (V, E)$ comprise a set of vertices, $V = \{v : v = 1, \cdots, L\}$, where $L = 2(n_l - 1)$ for unrooted and $L = 2n_l - 1$ for rooted trees,

and a set of edges, $E = \{(i,j) : i,j \in V, i < j\}$, where i and j are the vertices at the ends of the edge (i,j). Vertices are ordered such that for each edge (i,j), j is the parent (i.e., closer to the root) of i. For unrooted trees, the vertex chosen to represent the root is arbitrary and may even be a vertex of degree 1 (i.e., a leaf vertex).

The procedure for generating the topology requires the specification of a focal tree topology, T^*, and split probabilities for every split in T^*. This collection of splits will be denoted $\mathcal{S}(T^*)$, and the split probability for split s will be denoted $p(s)$.

In practice, T^* will be a fully resolved tree topology with high posterior probability based on the pilot run — the MAP (maximum a posteriori) tree, for example. If $0 < p(s) < 1$ for every split in T^*, then the procedure outlined below will specify a probability distribution over all n_t possible tree topologies. To guarantee this, we can base $p(s)$ on the frequency of split s in the pilot run. For example, if split s was sampled n_s times out of a total pilot run length n, and if s partitions the n_l taxa into two subsets of size n_1 and n_2 ($n_1 + n_2 = n_l$), then

$$n_s^* = \frac{(2n_1 - 3)!}{2^{n_1-2}(n_1 - 2)!} + \frac{(2n_2 - 3)!}{2^{n_2-2}(n_2 - 2)!},$$
$$p(s) = \frac{n_s + (n_s^*/n_t)\epsilon}{n + \epsilon}, \tag{5.1}$$

where ϵ is a small fraction of the MCMC sample size (e.g., $\epsilon = 0.01n$) and n_s^*/n_t is the induced prior on split s given a prior distribution that places equal weight on all possible unrooted tree topologies.

5.3.1.1 Tree simulation

Algorithm TopoGen describes how to generate a tree topology using the focal tree T^* and split probabilities for every split in $\mathcal{S}(T^*)$. The algorithm chooses which of the splits in $\mathcal{S}(T^*)$ to add. Because the splits are mutually compatible, they can be combined into a (possibly unresolved) tree. Finally, all polytomies are resolved by drawing a topology for their subtree, and rejecting any resolution of that portion of the tree that includes splits found in $\mathcal{S}(T^*)$.

Figure 5.1 illustrates simulation of a 6-taxon tree topology using the TopoGen algorithm. In the example illustrated, $S(T^*)$ comprises 3 splits, 2 of which, by chance, are included in $S(T_u)$, leaving T_u with a single polytomy of degree 4 at node v. The 4-taxon unrooted tree whose tips are vertices in $\mathcal{A}(v)$ has 3 possible resolutions, which are shown along the bottom of Figure 5.1. One of these possible resolutions (left) is invalid because it contains a split in T^*. One of the 2 remaining trees would be returned by the TopoGen algorithm with probability 0.5.

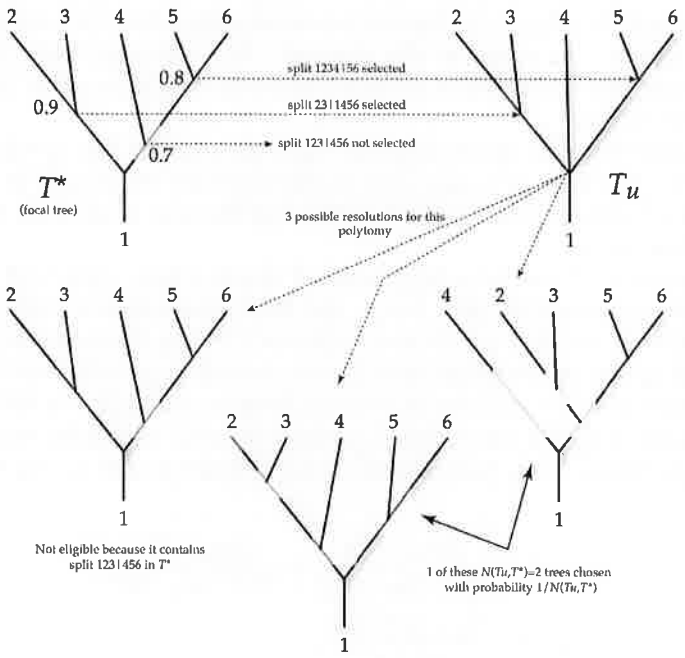

Figure 5.1: An example application of the TopoGen algorithm for a 6-taxon case.

5.3.1.2 Tree topology reference distribution

Let $\mathcal{S}(T_u)$ be the set of splits displayed by T that are members of the set $\mathcal{S}(T^*)$. Let T_u denote the (potentially nonbinary) tree that displays only those splits that are in $\mathcal{S}(T_u)$ (see TopoGen line 5). The probability of the tree topology component of the reference distribution is

$$\pi(T) = \Pr(T|p, \mathcal{S}(T^*)) = \Pr(T_u|p, \mathcal{S}(T^*)) \Pr(T|T_u, \mathcal{S}(T^*)). \tag{5.2}$$

The first term, $\Pr(T_u|p, \mathcal{S}(T^*))$, corresponds to lines 1–5 of the TopoGen algorithm, and equals the product of the probabilities of selecting each of the non-trivial splits found in T_u multiplied by the product of the probabilities of not selecting the other splits. For split s, let $I_{s\in\mathcal{S}(T_u)} = 1$ if $s \in \mathcal{S}(T_u)$ and 0 otherwise. Then, we have

$$\Pr(T_u|p, \mathcal{S}(T^*)) = \prod_{s\in\mathcal{S}(T^*)} p(s)^{I_{s\in\mathcal{S}(T_u)}} (1 - p(s))^{1-I_{s\in\mathcal{S}(T_u)}}. \tag{5.3}$$

From the fact that $\epsilon > 0$ and the definition of $p(s)$ in (5.1), $0 < p(s) < 1$ and therefore $0 < \Pr(T_u|f) < 1$. The independence of the probability terms in (5.3)

Algorithm 1 TopoGen: Generate a tree topology using split selection probabilities.

Require: $\mathcal{S}(T^*)$, a set of compatible splits, and function, p, that maps each split in $\mathcal{S}(T^*)$ to a probability that the split will be included in the tree.

1: $\mathcal{S}(T_u) \leftarrow \emptyset$
2: **for all** $s \in \mathcal{S}(T^*)$ **do**
3: With probability $p(s)$, add s to $\mathcal{S}(T_u)$
4: **end for**
5: Produce the tree T that displays the splits in $\mathcal{S}(T_u)$ (This tree, without the modifications below, will be referred to as T_u.)
6: **while** T is not fully resolved **do**
7: Randomly select a vertex, v, from T that is not fully resolved ($\deg(v) > 3$). Let $\mathcal{A}(v)$ be the set of vertices that are adjacent to v; i.e., $\mathcal{A}(v) = \{a : (i,j) \in E, a \in \{i,j\}, v \in \{i,j\}, a \neq v\}$.
8: Treat the vertices in $\mathcal{A}(v)$ as leaves and randomly choose tree topology, $T_{\mathcal{A}(v)}$, for them from a uniform distribution over all fully resolved unrooted trees with $|\mathcal{A}(v)|$ leaves (thus replacing the polytomy, v, with a set of vertices each of degree 3).
9: If any of the newly added splits (splits that correspond to internal edges in $T_{\mathcal{A}(v)}$) are in $\mathcal{S}(T^*)$, then collapse $T_{\mathcal{A}(v)}$ back into a polytomy (returning the tree to its original state before $T_{\mathcal{A}(v)}$ was added).
10: **end while**
11: **return** T as the simulated tree and $\mathcal{S}(T_u)$ as the set of splits from $\mathcal{S}(T^*)$ which are displayed by T.

reflects the fact that split selection in line 3 of the TopoGen algorithm does not depend in any way on splits already chosen to be included in set $\mathcal{S}(T_u)$.

The second term in $\pi(T)$, $\Pr(T|T_u, \mathcal{S}(T^*))$, corresponds to lines 6–10 of the TopoGen algorithm. Let $\mathcal{V}(T_u)$ denote the set of unresolved vertices in T_u:

$$\mathcal{V}(T_u) = \{v : v \in T_u, \deg(v) > 3\}.$$

Let $\mathcal{A}(v)$ denote the vertices that are adjacent to vertex v:

$$\mathcal{A}(v) = \{a : (i,j) \in E, a \in \{i,j\}, v \in \{i,j\}, a \neq v\}.$$

For each vertex v in $\mathcal{V}(T_u)$, we can easily identify the set of vertices $\mathcal{A}(v)$. For $x \in \mathcal{A}(v)$, let x^* be the corresponding vertex in T^*.

Let $T^*_{\mathcal{A}(v)}$ denote the tree that would be obtained from T^* by deleting vertices and edges such that $\mathcal{A}(v)$ is the leaf set of the new tree. The rejection step in TopoGen guarantees that the polytomy-breaking portion of the algorithm is equivalent to drawing from a discrete uniform distribution of all of the trees with leaf-set $\mathcal{A}(v)$ that share no splits with $T^*_{\mathcal{A}(v)}$. Bryant and Steel (2009) provide an algorithm for computing $q_s(T)$, the number of fully resolved unrooted trees that share exactly s splits with tree topology T. Using

their notation, the number of trees with leaf-set $\mathcal{A}(v)$ that share zero splits with $T^*_{\mathcal{A}(v)}$ is $q_0(T^*_{\mathcal{A}(v)})$. The algorithms presented below as "Preprocessing For Count Max Diff" and "Count Max Diff" were devised by one of us (DB); they provide a more efficient method of calculating $q_0(T^*_{\mathcal{A}(v)})$.

Algorithm 2 Preprocessing for Count Max Diff: Preprocessing steps for algorithm "Count Max Diff."

Require: T is a binary tree with n leaves rooted at an internal vertex v_0.
{*Pre-processing*}
$b[0] \leftarrow 1$
for $k = 1, 2, \ldots, (n-3)$ **do**
  $b[k] \leftarrow (2k+1)b[k-1]$ {$b[k]$ = number of binary trees on $k+3$ leaves}
end for
for v in a post-order traversal of T **do**
  if v is a leaf **then**
    $n[v] \leftarrow 0$
  else
    let v_1, v_2 be the children of v
    $n[v] \leftarrow n[v_1] + n[v_2]+$ number of children of v that are internal
    {$n[v]$ is the number of internal edges below v}
  end if
end for

$N(T_u, T^*)$ is the number of trees that are resolutions of T_u and contain no splits in $\mathcal{S}(T^*)$ other than the splits $\mathcal{S}(T_u)$ that can be found by considering the products of the number of all relevant resolutions around the set of unresolved vertices, $\mathcal{V}(T_u)$:

$$N(T_u, T^*) = \prod_{i \in \mathcal{V}(T_u)} q_0(T^*_{\mathcal{A}(i)}). \tag{5.4}$$

Because the TopoGen algorithm chooses uniformly from these trees,

$$\Pr(T|T_u, \mathcal{S}(T^*)) = \frac{1}{N(T_u, T^*)}. \tag{5.5}$$

Substituting into (5.2) yields:

$$\Pr(T|f, \mathcal{S}(T^*)) = \frac{\prod_{s \in \mathcal{S}(T^*)} p(s)^{I_{s \in \mathcal{S}(T_u)}} (1-p(s))^{1-I_{s \in \mathcal{S}(T_u)}}}{N(T_u, T^*)}. \tag{5.6}$$

To evaluate the probability of an arbitrary tree, T, (a tree for which T_u is not available beforehand), let T_u equal the strict consensus of T and T^*.

An example calculation of the probability of a 6-taxon tree T is illustrated in Figure 5.2. Tree T_u represents the strict consensus of T and T^*, and tree $T^*_{\mathcal{A}(v)}$ equals T^* with all vertices pruned except those corresponding to vertices in $\mathcal{A}(v)$ and rooted at internal node v. Results of applying the algorithms

Algorithm 3 Count Max Diff: Count the number of binary trees at the maximum RF distance from a focal tree

for vertex v in a post-order traversal of T **do**
  if v is internal with no children that are internal **then**
    $f[v,0] \leftarrow 1$
  else if v has one child v_1 that is internal **then**
    $f[v,0] \leftarrow -\sum_{\ell=0}^{n[v_1]} f[v_1,\ell]$
    for $k = 1,2,\ldots n[v]$ **do**
      $f[v,k] \leftarrow (2k+1)f[v_1,k-1]$
    end for
  else if v has two children v_1, v_2 that are internal **then**
    $F_1 \leftarrow \sum_{\ell=0}^{n[v_1]} f[v_1,\ell]$
    $F_2 \leftarrow \sum_{\ell=0}^{n[v_2]} f[v_2,\ell]$
    $f[v,0] \leftarrow F_1 \times F_2$
    for $k = 1,2,\ldots n[v]$ **do**
      $f[v,k] \leftarrow 0$
    end for
    for $k = 1,2,\ldots (n[v_2]+1)$ **do**
      $f[v,k] \leftarrow f[v,k] - F_1 \times f[v_2,k-1] \times (2k+1)$
    end for
    for $k = 1,2,\ldots (n[v_1]+1)$ **do**
      $f[v,k] \leftarrow f[v,k] - F_2 \times f[v_1,k-1] \times (2k+1)$
    end for
    for $k_1 = 1,2,\ldots (n[v_1]+1)$ **do**
      for $k_2 = 1,2,\ldots (n[v_2]+1)$ **do**
        $k \leftarrow k_1 + k_2$
        $f[v,k] \leftarrow f[v,k] + f[v_1,k_1-1] \times f[v_2,k_2-1] \times \frac{b[k]}{b[k_1-1] \times b[k_2-1]}$
      end for
    end for
  end if
end for
return $\left|\sum_{k=0}^{n[v_0]} f[v_0,k]\right|$

"Preprocessing For Count Max Diff" and "Count Max Diff" are shown beside internal nodes of $T^*_{\mathcal{A}(v)}$, and the calculation of the probability of tree T is given at the bottom.

5.3.2 Edge length reference distribution

An edge length reference distribution can be constructed using a sample from the MCMC pilot run. Let $\mathcal{B} = \{s : n_s > N_{\min}\}$ be the set of splits having sample size at least $N_{\min}$ in the pilot run. The probability density of the edge

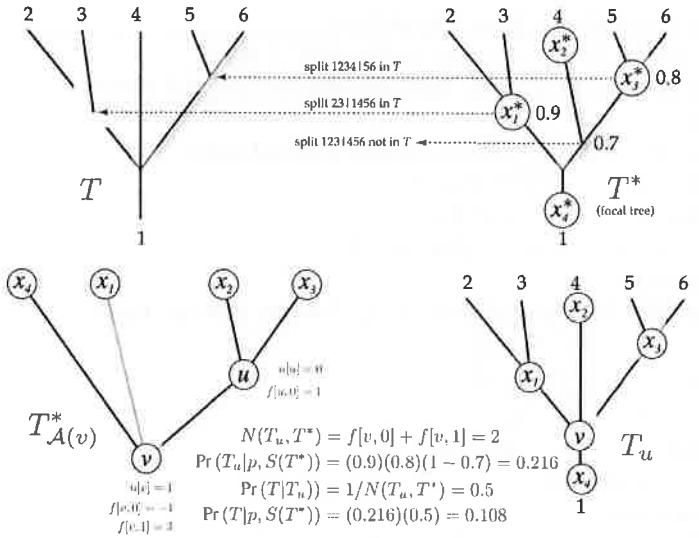

Figure 5.2: Calculation of $\Pr(T|p, S(T^*))$ for a 6-taxon example.

length ν_s corresponding to split s is

$$f(\nu_s) = \begin{cases} \text{Gamma}(a_s, b_s) & \text{if } s \in \mathcal{B}, \\ \text{Gamma}(a., b.) & \text{otherwise,} \end{cases}$$

where a_s and b_s are split-specific shape and scale parameters of a Gamma distribution fit to the marginal posterior distribution of ν_s from the pilot run, and $a.$ and $b.$ are the shape and scale parameters of a Gamma distribution fit to the marginal posterior distribution of edge lengths $\nu.$ associated with any split not in $\mathcal{B}$. (The choice of a Gamma distribution here is arbitrary, and could be replaced by a Lognormal distribution, or any other univariate probability distribution with support $(0, \infty)$.) The edge length reference distribution $f(\boldsymbol{\nu}_T|T)$ may now be defined as:

$$f(\boldsymbol{\nu}_T|T) = \prod_{s \in \mathcal{S}(T)} f(\nu_s).$$

The value N_{min} should be chosen large enough to provide reliable estimates of μ_s and σ_s.

5.3.3 Comparison with CCD methods

Larget (2013) and Höhna and Drummond (2011) use conditional clade distributions (CCDs) to provide approximations to marginal posterior distributions

of tree topologies. Both approaches use a preliminary sample from the posterior distribution to estimate CCDs and, from those, allow estimation of the marginal posterior probability of an arbitrary tree topology. Larget's method improves upon Höhna and Drummond in being more accurate and not requiring normalization. The reference distribution reported here differs from both of these CCD methods in allowing simulation and calculation of the probability of tree topologies having conditional clade relationships not sampled in the preliminary pilot run. Our method is far less accurate than Larget's approach in general, but reference distributions used for stepping-stone can provide efficiency for marginal likelihood estimation even if not providing the most accurate approximation to the posterior distribution of trees. Nevertheless, development of an approach based on Larget (2013) that allows construction of an irreducible Markov chain is well worthy of future effort.

5.4 Example

Lewis and Trainor (2011) obtained *rbc*L chloroplast DNA sequences of 6 green algae in the genus *Protosiphon* and two closely related genera in order to identify the lone surviving green alga in soil kept dry for 43 years. The phylogeny of these 6 taxa provides a good test of variable-topology marginal likelihood estimation methods because the posterior distribution is not dominated by a single tree topology.

5.4.1 Model details

A general time reversible (GTR) substitution model (Lanave et al., 1984; Tavaré, 1986) allowing invariable sites (Reeves, 1992) and discrete-gamma among-site rate heterogeneity (Yang, 1994b) was used to model evolution of DNA sequences along edges of the tree. The GTR model has 8 free parameters (3 equilibrium relative nucleotide frequencies and 5 exchangeability parameters), with 2 more parameters added to model among-site rate heterogeneity (proportion of invariable sites and discrete gamma shape parameter). Restrictions were placed on these 10 parameters to create a total of 12 different models. An additional 12 models resulted from partitioning the data by codon position. Noninformative Dirichlet distributions were used as priors for relative nucleotide frequencies and exchangeabilities, with a noninformative Beta distribution used for the proportion of invariable sites. An Exponential(1) distribution was used as the prior for the discrete gamma shape, and each edge length was assigned an independent Exponential prior distribution with mean 0.1. The prior probability for each tree topology was 1/105, where applicable (i.e., discrete uniform distribution over all possible unrooted tree topologies). For partitioned models, a noninformative relative rate distribution (eq. 3 in Fan

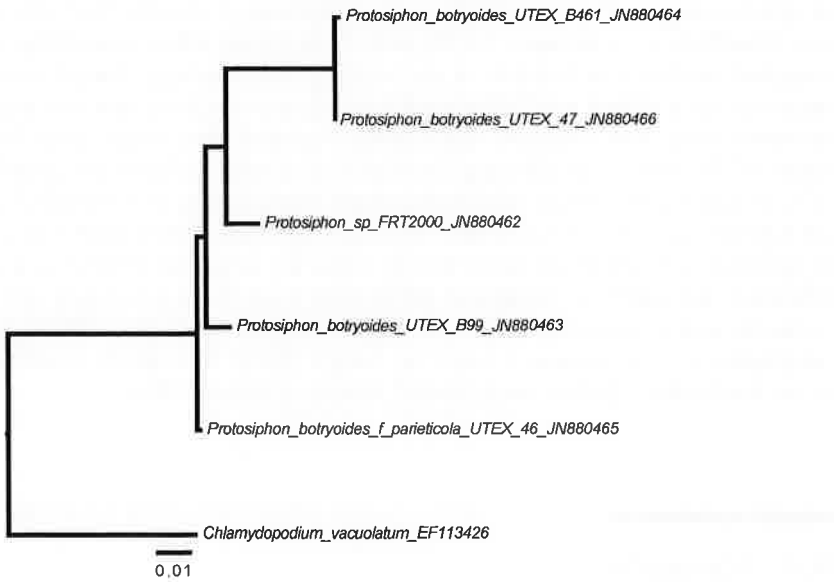

Figure 5.3: Tree topology with maximum marginal likelihood.

et al., 2011) was used as the prior for subset relative rates. In this study, all parameters of the relative rate distribution equaled 1, resulting in the constant relative rate density $2p_1p_2 = 2/9$ (where $p_1 = 1/3$ and $p_2 = 1/3$ are the proportions of sites in the 1st and 2nd position subsets, respectively).

5.4.2 Brute-force approach

It is not possible to simulate sequence data with a known marginal likelihood, and it is not possible to compute the marginal likelihood analytically for phylogenetic datasets of any reasonable size or complexity, so we must resort to other approaches to test the accuracy of particular estimation methods. Because estimating the marginal likelihood with GSS has been demonstrated to be accurate when the tree topology is fixed, one way to test the method proposed here is to estimate the total marginal likelihood from individual fixed tree results. The total marginal likelihood for this 6-taxon example may be written

$$p(\mathbf{y}) = \sum_{i=1}^{105} p(\mathbf{y}|T_i)p(T_i),$$

where T_i is the ith tree topology (out of the 105 tree topologies possible for 6 taxa), $p(T_i) = 1/105$ (discrete uniform prior distribution), and

$$p(\mathbf{y}|T_i) = \int p(\mathbf{y}|T_i, \theta)p(\theta|T_i)d\theta$$

is the conditional marginal likelihood given tree T_i.

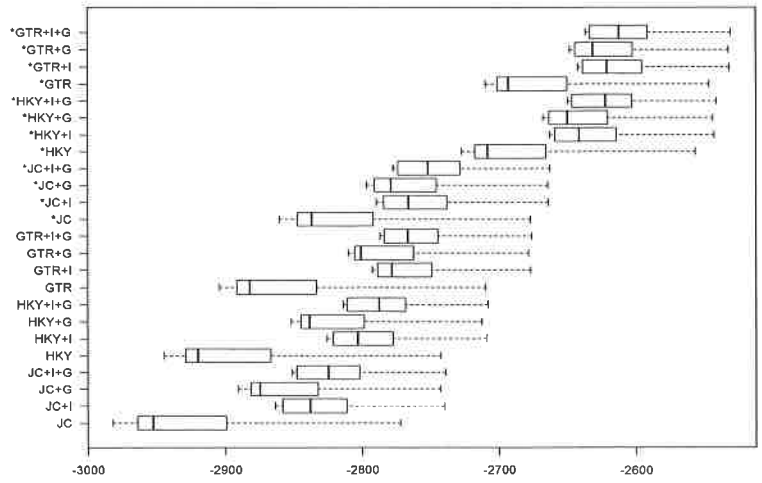

Figure 5.4: Box plots showing mean, 25%, and 75% quantiles and extremes for the 105 log marginal likelihoods estimated for each of the 24 models.

Figure 5.4 provides box plots of $\log p(\mathbf{y}|T_i)$ across the 105 tree topologies for each model. Partitioning sites into three subsets according to codon position provides the most dramatic increase in model fit; however, adding rate heterogeneity (+I, +G, or +I+G) to any unpartitioned model also substantially improves its fit relative to the base model (JC, HKY, or GTR). Despite differences in fit, the tree topology in Figure 5.3 (and Fig. 4 of Lewis and Trainor, 2011) was best according to log marginal likelihood for every one of the 24 models tested.

With estimates of all 105 conditional marginal likelihoods, it is possible to very accurately estimate the tree topology posterior distribution:

$$p(T|\mathbf{y}) = \frac{p(\mathbf{y}|T)p(T)}{p(\mathbf{y})}.$$

Figure 5.5 is a bar plot of the tree topology posterior distribution ordered left to right by Robinson-Foulds symmetric difference (RFSD) distance (Robinson and Foulds, 1981) from the best tree. The best tree has posterior probability 0.584, the 6 tree topologies 1 nearest-neighbor interchange (NNI) away from the best tree (RFSD distance = 2) collectively contributed 0.304 posterior probability, the 24 tree topologies 2 NNI swaps away (RFSD distance = 4) collectively account for 0.112 posterior probability, and none of the 74 tree topologies with the maximum RFSD distance (6) from the best tree topology contributed appreciably to the posterior.

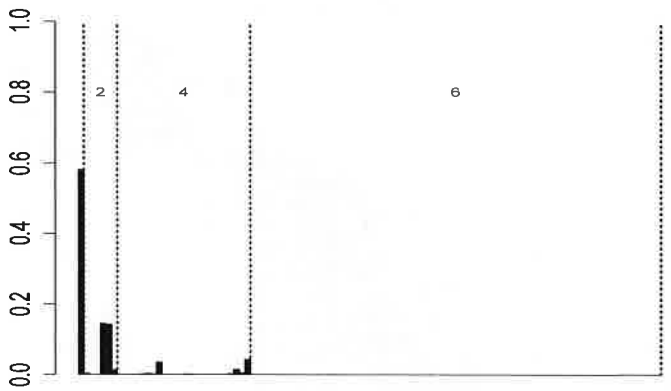

Figure 5.5: Bar plot of the tree topology posterior distribution. The 105-tree topologies are grouped by their Robinson-Foulds symmetric difference (RFSD) distance from the best tree (first bar on left). Groups are separated by vertical dotted lines, and number indicate the RFSD distance for each group.

5.4.3 GSS performance

The variable-topology GSS method performed well over all 24 models tested. Table 5.1 ranks models from best to worst according to their (brute-force) marginal likelihood estimates. For each model, the GSS estimate is given alongside the brute-force estimate, and the column labeled Δ is the difference (GSS estimate minus brute-force estimate). Most absolute differences are less than 0.1 log unit.

Not surprisingly, the accuracy of both the GSS and brute-force approaches depends on the quality of the samples obtained during the stepping-stone MCMC simulation, and both can deviate by several log units from the correct value if care is not taken to ensure that autocorrelation is minimized for all parameters and for all β values. One benefit of using GSS over SS is that in GSS the MCMC simulation crawls between the posterior and the reference distribution, which is similar to the posterior in many respects. The expectation is that an MCMC simulator tuned to the posterior still mixes well when exploring the reference distribution. In contrast, SS crawls from the posterior to the prior. Given that the prior is typically much less informative than the posterior, considerable adjustments must be made to the tuning of the MCMC simulator as the analysis proceeds. Proposals that are bold for the posterior represent tiny steps as the MCMC nears the prior. Despite our expectations, we found that some adjustment to tuning was necessary even for GSS. The reference distribution used is essentially the prior fit to samples drawn from a pilot study of the posterior. This approach matches means and

Table 5.1: Comparison of marginal likelihoods for 24 models estimated using the generalized stepping-stone (GSS) and brute-force approaches. Δ is the difference between brute-force and GSS marginal likelihood estimates. "From best" is the difference in the brute-force estimate from that of the best model (GTR+I+G*). Models indicated by an asterisk (*) partitioned sites by codon position. All other models used a single model for all sites.

Model	GSS	Brute force	Δ	From best
*GTR+I+G	−2534.57	−2534.66	0.09	0.00
*GTR+I	−2535.59	−2535.57	−0.02	0.91
*GTR+G	−2536.69	−2536.75	0.06	2.09
*HKY+I+G	−2544.49	−2545.04	0.55	10.38
*HKY+I	−2546.78	−2546.75	−0.03	12.09
*HKY+G	−2548.20	−2548.16	−0.04	13.50
*GTR	−2551.09	−2551.10	0.01	16.44
*HKY	−2561.10	−2561.13	0.03	26.47
*JC+I+G	−2667.09	−2667.19	0.10	132.53
*JC+I	−2668.35	−2668.38	0.03	133.72
*JC+G	−2668.94	−2668.99	0.05	134.33
GTR+I+G	−2680.21	−2680.29	0.08	145.63
GTR+I	−2681.01	−2681.00	−0.01	146.34
*JC	−2681.83	−2681.79	−0.04	147.13
GTR+G	−2682.38	−2682.73	0.02	148.07
HKY+I+G	−2712.18	−2712.30	0.12	177.64
HKY+I	−2713.28	−2713.31	0.03	178.65
GTR	−2714.22	−2714.20	−0.02	179.54
HKY+G	−2716.83	−2716.87	0.04	182.21
JC+I+G	−2743.19	−2743.56	0.37	208.90
JC+I	−2744.48	−2744.59	0.11	209.93
HKY	−2747.01	−2746.99	−0.02	212.33
JC+G	−2747.44	−2747.44	0.00	212.78
JC	−2776.58	−2776.52	−0.06	241.86

variances from the posterior in constructing the reference distribution, but fails to capture correlations among parameters. Such correlations are particularly strong between certain edge length parameters, and between edge lengths and rate heterogeneity parameters. As a result, many values sampled from the joint reference distribution may be quite improbable with respect to the posterior, and we found that increasing the boldness of some proposals as a function of β helped keep autocorrelation low.

5.5 Summary

This chapter describes a method for generalizing the generalized stepping-stone (GSS) method to accommodate varying tree topology. The GSS method as originally described (Fan et al., 2011) was designed for estimating the marginal likelihood under a fixed tree topology. Systematists, in particular, use Bayesian methods expressly to estimate the tree topology, and would therefore prefer to base model comparison on $p(\mathbf{y})$ rather than $p(\mathbf{y}|T)$ for some fixed tree topology T. The greater efficiency of the GSS method compared to SS (Xie et al., 2011) is its use of a reference distribution that is close (as measured by Kullback-Leibler divergence) to the posterior distribution. The SS method instead uses the prior distribution as the reference distribution, and the prior is usually quite different than the posterior, usually displaying a much greater variance.

The primary contribution of this chapter is a proposed reference distribution for trees (topology and edge lengths) that assigns high probability to topologies containing splits deemed important by the posterior. In addition to tree topology, edge length distributions are maintained separately for high (posterior)-probability splits to improve the match between reference distribution and posterior distribution. The proposed reference distribution for trees has a known normalizing constant, which is a requirement for any reference distribution used within the context of GSS.

Marginal likelihoods were estimated for 24 models for a 6-taxon example using the variable-topology GSS. The results compared well to marginal likelihoods estimated using a brute-force approach that involved estimating the marginal likelihood separately for each of the 105 possible unrooted tree topologies, then combining these to give the total marginal likelihood. The method proposed here applies to binary unrooted trees. Adapting the method to rooted trees and analyses involving polytomous trees (see Lewis et al., 2005) will require further work.

The GSS method described here is implemented in Phycas version 2.0 (freely available at `phycas.org`). Python and bash scripts for performing all analyses reported here are available in the supplementary materials for the book.

5.6 Funding

This material is based upon work supported by the National Science Foundation under grant numbers DEB-0732920, DEB-1208393 (MTH), and DEB-1036448 (POL). Any opinions, findings, and conclusions or recommendations expressed

in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgments

This work made extensive use of the computer cluster maintained by the Bioinformatics Facility (Biotechnology*Bioservices Center), University of Connecticut (http://bioinformatics.uconn.edu/).