DOI: 10.1534/genetics.103.025692

Stochastic Models for Horizontal Gene Transfer: Taking a Random Walk Through Tree Space

Marc A. Suchard[1]

Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, California 90095-1766

Manuscript received December 11, 2003
Accepted for publication February 1, 2005

ABSTRACT

Horizontal gene transfer (HGT) plays a critical role in evolution across all domains of life with important biological and medical implications. I propose a simple class of stochastic models to examine HGT using multiple orthologous gene alignments. The models function in a hierarchical phylogenetic framework. The top level of the hierarchy is based on a random walk process in "tree space" that allows for the development of a joint probabilistic distribution over multiple gene trees and an unknown, but estimable species tree. I consider two general forms of random walks. The first form is derived from the subtree prune and regraft (SPR) operator that mirrors the observed effects that HGT has on inferred trees. The second form is based on walks over complete graphs and offers numerically tractable solutions for an increasing number of taxa. The bottom level of the hierarchy utilizes standard phylogenetic models to reconstruct gene trees given multiple gene alignments conditional on the random walk process. I develop a well-mixing Markov chain Monte Carlo algorithm to fit the models in a Bayesian framework. I demonstrate the flexibility of these stochastic models to test competing ideas about HGT by examining the complexity hypothesis. Using 144 orthologous gene alignments from six prokaryotes previously collected and analyzed, Bayesian model selection finds support for (1) the SPR model over the alternative form, (2) the 16S rRNA reconstruction as the most likely species tree, and (3) increased HGT of operational genes compared to informational genes.

TRADITIONAL views of molecular evolution hold that genetic material mutates slowly over time as it is passed in a vertical fashion from parent to progeny. Molecular phylogenetics then aims to reconstruct this history of inheritance of genetic sequence data from contemporary organisms into a tree-like structure. However, belief in a single tree, mandated by vertical transmission, for all genetic material is changing. Evolutionary biologists increasingly recognize the horizontal transmission of genetic material between distantly related organisms as an important mechanism of evolution (Syvanen 1994; Lawrence 1999; Jain *et al.* 2002).

The process of horizontal (or lateral) gene transfer (HGT) plays a critical role across all domains of life and in particular among prokaryotes (Jain *et al.* 1999; Koonin *et al.* 2001). For example, many prokaryotes are agile at quickly adapting to new environments. Often, this ability stems from the acquisition of new genes through HGT rather than through random mutation (Lawrence 1999). At least three mechanisms promote HGT in prokaryotes (Jain *et al.* 2002). These include: (1) transformation in which free DNA sequences are absorbed from the environment, (2) conjugation between two different prokaryotic species, and (3) transduction of genetic material through viruses. Finally, HGT also has medical importance (Brown 2003). In the field of infectious diseases, HGT among bacterial pathogens of antibiotic resistance genes has greatly contributed to the emergence of multidrug-resistant bacteria in clinical settings (Leverstein-van Hall *et al.* 2002). In the field of oncology, HGT may also affect tumor progression; Bergsmedh *et al.* (2001) show that eukaryotic cells can transfer active oncogenes.

Three general methods have been employed to examine HGT. The first focuses on single genomes and identifies genes suspected to have been imported through HGT by examining variation in nucleotide base composition and codon usage patterns (Lawrence and Ochman, 1997). The latter two methods are comparative studies across species. One uses similarity approaches based on gene content to identify HGT (Ragan 2001) and to propose average genome or species-level trees (Snel *et al.* 1999), while the alternative method endorses phylogenetic reconstruction using orthologous genes (Jain *et al.* 1999). Base composition and codon bias studies may perform poorly when compared to phylogenetic methods (Koski *et al.* 2001). Further, phylogenetic methods offer at least one advantage over similarity-based approaches. The reconstructed phylogenies have direct biological interpretability as descriptions of the underly-

[1]*Address for correspondence:* Department of Biomathematics, David Geffen School of Medicine, UCLA, 650 Charles Young Dr., Box 951766, Los Angeles, CA 90095-1766. E-mail: msuchard@ucla.edu

ing evolutionary histories of the different genes (Doolittle 1999). If a reconstructed gene tree differs from the assumed phylogeny of the species being studied, then HGT is offered as a possible explanation (Syvanen 1994). One intrinsic difficulty is that the true species tree is often itself unknown. Therefore, it is necessary to either fix the species tree to equal the inferred gene tree for a specially chosen gene, *e.g.*, the 16S rRNA tree (Woese 2000), or simultaneously estimate the species tree and gene trees given a biologically plausible model relating them. As a first step, several research groups have attacked the inverse problem of reconstructing a species tree given gene trees subject to HGT. Most notable are the parsimony-based reconciled tree work by Page and colleagues (*e.g.*, Page 2000) and the algorithmic work of Mirkin *et al.* (2003).

I propose a simple class of stochastic models for HGT that enable the simultaneous estimation of the underlying species tree relating a group of organisms and the gene trees subject to HGT for a set of orthologous gene alignments. These HGT models function in a hierarchical manner (Suchard *et al.* 2003a) in which standard Bayesian phylogenetic approaches (*e.g.*, Sinsheimer *et al.* 1996; Yang and Rannala 1997; Mau *et al.* 1999; Li *et al.* 2000; Huelsenbeck *et al.* 2001) are used to reconstruct each gene tree from its corresponding gene alignment. Simultaneous to the reconstructions, the HGT models impose a second probabilistic distribution over the gene trees (Maddison 1997). This hierarchical distribution describes the gene trees likelihoods given an unknown species tree and an unknown number of HGT events leading from that species tree to each gene tree. The model is fit in a Bayesian framework that naturally handles uncertainty in discrete parameters such as all the trees and the number of HGT events and compares various models using Bayes factors (Suchard *et al.* 2001). Stochastic models fit in statistical frameworks offer several advantages over parsimony approaches. First, parsimony may underestimate the number of HGT events linking the species tree to the gene trees. This consequence is similarly seen in parsimonious reconstructions of the tree themselves, in which the number of nucleotide substitutions is underestimated. Second, it is easier in a statistical framework to include measures of uncertainty and these levels may be high in the inferred gene trees given the sparse data from which they are reconstructed.

One additional advantage of building stochastic models for HGT is the ability to compare competing models and to incorporate possible differences in the stochastic processes across genes, while assessing the significance of these differences in a formal statistical framework. As one example of possible differences across genes, Jain *et al.* (1999) propose the complexity hypothesis. Under this hypothesis, genes are divided into one of two classes, informational or operational genes. Between classes, the rates of HGT differ. It is suspected that rates are higher for operational genes than for informational genes. This hypothesis and others can be tested by integrating over all possible species trees and gene trees weighed by their posterior probabilities. This Bayesian model-averaging approach reduces the possible bias inherent in selecting a specific species tree, minimizes underestimation of the uncertainty associated with the hypotheses (Taylor *et al.* 1996), and eliminates the need for *ad hoc* analyses. Formal comparison of different models for HGT will help gather further insight into the underlying biological processes.

MODEL

Within-gene reconstruction model: I begin with a hierarchical framework for phylogenetic reconstruction using molecular sequence data $\mathbf{Y}$ (Suchard *et al.* 2003a). Data $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ consist of K naturally disjoint partitions. Partition data $\mathbf{Y}_k$ for $k = 1, \ldots, K$ represent the aligned DNA sequences of length L_k from one specific gene per partition, sequenced from the same N taxa across all partitions. A hierarchical phylogenetic model enables the pooling of information across gene partitions to improve estimate precision in individual partitions, while permitting estimation and testing of tendencies in across-partition quantities. For HGT, such across-partition quantities include: (1) an overall species tree, (2) appropriate stochastic models from which to construct a probability distribution over individual gene trees given the species tree, and (3) the stochastic model parameters that may vary between different classes of genes.

To utilize standard Bayesian models for phylogenetic reconstruction (*e.g.*, Sinsheimer *et al.* 1996; Yang and Rannala 1997; Mau *et al.* 1999; Li *et al.* 2000) within a gene partition, data $\mathbf{Y}_k$ further divide into ordered homologous sites $\mathbf{Y}_{kl}$ for $l = 1, \ldots, L_k$. Site data $\mathbf{Y}_{kl} = (Y_{kl1}, \ldots, Y_{klN})^t$ contain one nucleotide from each taxon, such that $Y_{kln} \in$ (A, G, C, T) or their ambiguous wildcards for $n = 1, \ldots, N$. I assume that sites within a partition are independent and identically distributed, and the likelihood of observing $\mathbf{Y}_{kl}$ is given by a multinomial distribution over the 4^N possible outcomes with ambiguous nucleotides being integrated over their possible realizations. The multinomial outcome probabilities become functions of an unknown tree τ_k that describes the relatedness of the N taxa, branch lengths $\mathbf{t}_k = (t_{k1}, \ldots, t_{kB})$, and a model to describe nucleotide mutation along these branches, all within partition k.

I elect for a reversible, continuous-time Markov chain (CTMC) model for nucleotide substitution (Felsenstein 1981) popularized by Tamura and Nei (1993) (TN93). The TN93 model is further parameterized by two transition:transversion rate ratios, α_k between purines A and G and γ_k between pyrimidines C and T, and the stationary distribution of the underlying Markov chain $\boldsymbol{\pi}_k = (\pi_{kA}, \pi_{kG}, \pi_{kC}, \pi_{kT})$. The final scale parameter in the TN93 model is fixed such that branch lengths measure the expected number of nucleotide substitu-

tions between the nodes in τ_k that the branch connects. Because I assume a reversible model for nucleotide substitution and make no clock-like restrictions on branch lengths, the root of each tree is unidentifiable (Felsenstein 1981). As a consequence, the descriptions of all trees to follow are unrooted with $N - 2$ internal nodes and $B = 2N - 3$ branches.

Across-gene hierarchical model: Following the hierarchical framework of Suchard *et al.* (2003a), I take branch lengths $\mathbf{t}_k$ as exponentially distributed with unknown expected divergence μ_k within partition k and model

$$\begin{pmatrix} \log \alpha_k \\ \log \gamma_k \\ \log \mu_k \end{pmatrix} \sim \text{Normal}(\mathbf{V}, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{\pi}_k \sim \text{Dirichlet}(N_\Pi \times \boldsymbol{\Pi}), \tag{1}$$

where $\mathbf{V} = (A, G, M)^t$ and $\boldsymbol{\Pi} = (\Pi_A, \Pi_G, \Pi_C, \Pi_T)$ are unknown across-partition-level expectations, variance-covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma^2_\alpha, \sigma^2_\gamma, \sigma^2_\mu)$ has diagonal form, and σ^{-2}_α, σ^{-2}_γ, σ^{-2}_μ, and N_Π are unknown across-partition-level measures of precision. Leaving $\mathbf{V}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Pi}$, and N_Π as unknowns specified only by hyperprior distributions and estimating these parameters simultaneously with the within-partition-level continuous parameters, α_k, γ_k, and μ_k for all k, enables the borrowing of strength of information from one partition by another, producing more precise within-partition-level estimates. I assume conjugate (when possible) and flat or noninformative hyperpriors on these across-partition-level parameters, as discussed in Suchard *et al.* (2003a). While the development of hierarchical priors over the continuous within-partition-level parameters has been straightforward, constructing a hierarchical prior over gene trees τ_k that incorporates the stochastic nature of HGT is more involved. This is illustrated in the next section.

Horizontal gene transfer models: To build a stochastic model for HGT, I first present a formal description of the set of all possible N-taxon trees, commonly referred to as "tree space" (Billera *et al.* 2001), as a mathematical graph and then discuss several possible random walks (D. Aldous and J. Fill, unpublished results) on this graph that mirror the observed effects of HGT.

There exist $M = (2N - 5)!/2^{N-3}(N - 3)!$ possible trees relating N extant taxa (Felsenstein 1981). On the basis of these M trees, I construct a graph $\mathcal{G} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ with vertex set $\boldsymbol{\mathcal{V}}$ and edge set $\boldsymbol{\mathcal{E}}$. Each tree represents a different vertex, or node, in the graph, such that the size of the vertex set $|\boldsymbol{\mathcal{V}}| = M$. An edge $uv \in \boldsymbol{\mathcal{E}}$ of a graph describes a direct connection between two of the graph's vertices $u, v \in \boldsymbol{\mathcal{V}}$. The number of edges emanating from a single vertex v defines its degree $d(v)$. Two vertices that are joined together by a single edge are called adjacent. Restricting attention to simple graphs in which pairs of vertices may be connected to each other only by a single edge and no vertex is connected to itself by a looping edge, a single vertex v from graph $\mathcal{G}$ may be adjacent from as few as zero to as many as $M - 1$ other vertices. The set of all vertices adjacent to v are its neighborhood $\Gamma(v)$ and the size of this neighborhood $|\Gamma(v)| = d(v)$. The specification of a neighborhood for each vertex completes the description of $\mathcal{G}$, and many choices are available.

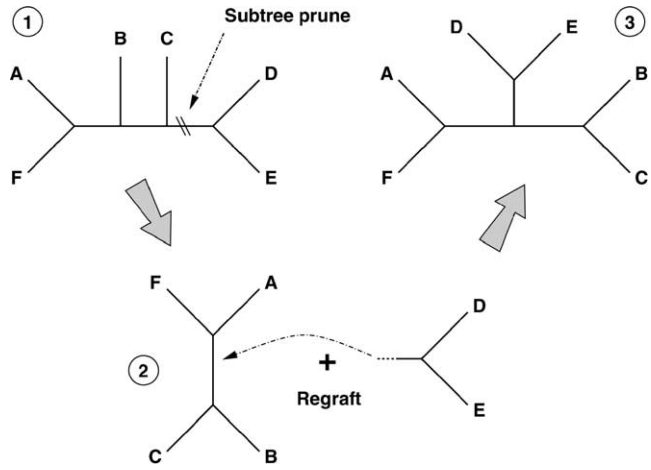

Figure 1.—Subtree prune and regraft operator applied to a six-taxon tree. (1) Operator selects and cuts any branch in the initial tree, pruning away a subtree. (2) Operator regrafts subtree by selecting and subdividing a preexisting branch in the remaining tree. (3) Resultant tree for this realization.

Subtree-prune-regraft-based model: One approach to defining neighborhoods for each possible tree stems from subtree transfer operations (Allen and Steel 2001). Subtree transfer operators act on trees producing local rearrangements. Applying a subtree transfer operator to one tree τ results in the creation of one of several possible new topologies that differs from τ by an extent dependent on the operator. The collection of all trees one operation away from $\tau = v$ becomes its neighborhood $\Gamma(v)$ under that operator. Nearest-neighbor interchange (Robinson 1971), tree bisection and reconnection (Swofford *et al.* 1996), and subtree prune and regraft (SPR) (Hein 1990, 1993) are three examples. In light of the goals of this article, SPR offers an advantage over the former two operators because of its potential biological interpretation. Applying the SPR operator to $\tau = v$ with its resultant drawn from $\Gamma_{\text{SPR}}(v)$ mirrors the differences observed between a species tree and an individual gene tree affected by one HGT (or recombination) event (Hein 1990, 1993; Jain *et al.* 1999; Allen and Steel 2001).

Figure 1 illustrates one realization of the SPR operator applied to a six-taxon tree. The operator works in two steps. The first step selects and cuts any branch in the initial tree, τ_{initial}. Cutting the branch prunes away a subtree, τ_{subtree}. This subtree then regrafts itself using

the same cut branch to a new internal node obtained by subdividing a preexisting branch in $\tau_{\text{initial}} - \tau_{\text{subtree}}$.

Several important properties about the graph $\mathcal{G}_{\text{SPR}}$ induced by the SPR operator have been previously studied. First, $\mathcal{G}_{\text{SPR}}$ is regular, implying that every vertex $v \in \mathcal{V}_{\text{SPR}}$ possesses the same degree $d(v) = 2(N-3)(2N-7)$ and, hence, neighborhood size (Allen and Steel 2001). Also, $\mathcal{G}_{\text{SPR}}$ is connected, meaning that a sequence of consecutive edges (a path) exists, connecting every pair of vertices in $\mathcal{G}_{\text{SPR}}$ (Robinson 1971; Allen and Steel 2001).

One straightforward stochastic process on any simple graph $\mathcal{G}$ is an unweighted random walk. A random walk on $\mathcal{G}$ proceeds from vertex to vertex along existing edges of the graph, generating a discrete-time Markov chain (DTMC), where the states of the chain are the visited vertices. As unweighted, the chain uniform randomly chooses its next vertex to visit from all neighbors of its current vertex. For this DTMC, the one-event transition probability matrix $\mathbf{A}$ has entries

$$(\mathbf{A})_{uv} = \begin{cases} \dfrac{1}{d(u)} & \text{if vertices } u \text{ and } v \text{ are adjacent or} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

defining the probability of u changing into v as a result of one random event. It should be noted that $\mathbf{A}$ is just the adjacency matrix of $\mathcal{G}$ rescaled to be a stochastic matrix [*i.e.*, $\Sigma_v(\mathbf{A})_{uv} = 1$].

On the basis of K random walks on the graph $\mathcal{G}_{\text{SPR}}$ induced by the SPR operator, I construct a hierarchical prior over the joint distribution of all gene trees τ_k. To accomplish this task, I assume:

An unknown species tree Υ exists.
The vertex representing Υ is the initial state of K Markov chains.
The Markov chains are conditionally independent given Υ and $\mathbf{A}$.
The vertex representing τ_k is the final state of the kth chain.
And each chain is of unknown length $0 \le E_k < \infty$.

Figure 2 depicts one set of the possible paths of $K = 4$ Markov chains starting at species tree Υ and ending at gene trees τ_k on a small portion of a representative graph. The lengths of paths E_k shown range from one to three. I illustrate no paths of length zero, but these realizations should be most likely. A parsimony-like analysis considering beginning and end points of the chains in Figure 2 would, for example, underestimate E_4 as zero instead of three.

Given the assumptions listed above, the probability of species tree Υ giving rise to gene tree τ_k after E_k HGT events is

$$q(\tau_k = v|\Upsilon = u, E_k) = (\mathbf{A}^{E_k})_{uv}. \tag{3}$$

To complete the hierarchical specification, I assign a prior distribution over Υ by letting

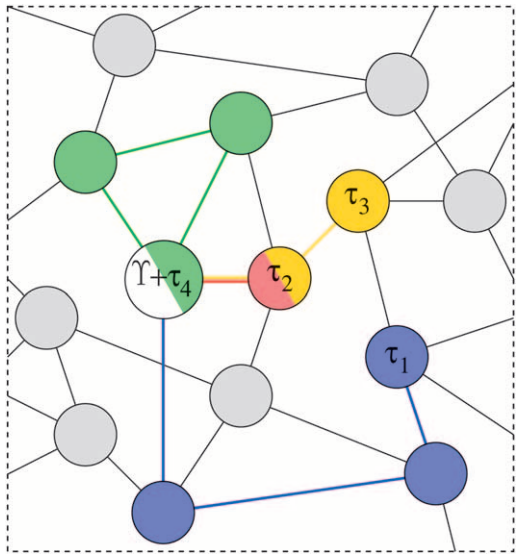

Figure 2.—One possible Markov chain realization on a simplified graph for the species tree Υ and four gene trees $\tau_1, \ldots, \tau_4$. All chains begin at the same vertex representing the species tree Υ (in white). The chain producing gene tree τ_1 has length $E_1 = 3$ (blue), the chain for τ_2 has length $E_2 = 1$ (red), the chain for τ_3 has length $E_3 = 2$ (yellow), and the chain for τ_4 has length $E_4 = 3$ (green). Note that this latter chain returns to its starting state; a parsimony-like analysis would estimate $E_4 = 0$. Not depicted are chains with actual length zero; these are most probable *a priori*.

$$\Upsilon \sim \text{Multinomial}(\mathbf{z}), \tag{4}$$

where $\mathbf{z} = (z_1, \ldots, z_M)$ are constants, the prior probabilities of the M possible N-taxon trees. When little or no information is available about Υ, one reasonable choice is $z_1 = \ldots = z_M = 1/M$; alternately, one may choose $\mathbf{z}$ such that the prior odds of competing hypotheses regarding Υ are one in a hypothesis-testing setting (Suchard *et al.* 2003a). A further choice is discussed later. I further assume a conditionally independent prior on all E_k,

$$E_k \sim \text{Poisson}(\Lambda_k), \tag{5}$$

where Λ_k is the expected number of HGT events for gene k and is a deterministic function of across-gene-level parameters. This prior is conjugate to (3), allowing all E_k to be integrated out of the model, improving sampling efficiency (Liu 1994),

$$q(\tau_k = v|\Upsilon = u, \Lambda_k) = \sum_{E_k=0}^{\infty} q(\tau_k = v|\Upsilon = u, E_k)\, q(E_k|\Lambda_k). \tag{6}$$

Letting $q(\tau_k = v|\Upsilon = u, \Lambda_k) = (\mathbf{P})_{uv}$, the multistep transition probability matrix,

$$\mathbf{P} = \sum_{E_k=0}^{\infty} \mathbf{A}^{E_k} q(E_k|\Lambda_k),$$

$$= \sum_{E_k=0}^{\infty} \mathbf{A}^{E_k} \exp(-\Lambda_k) \frac{\Lambda_k^{E_k}}{E_k!},$$

$$= \exp(-\Lambda_k)\exp(\Lambda_k \mathbf{A}),$$
$$= \exp\{\Lambda_k(\mathbf{A} - \mathbf{I})\} = \exp(\Lambda_k \mathbf{Q}), \quad (7)$$

where $\mathbf{I}$ is the $M \times M$ identity matrix and $\mathbf{Q} = \mathbf{P} - \mathbf{I}$ is the CTMC infinitesimal rate matrix representation of the HGT process. In this parameterization, Λ_k are scaled as the expected number of HGT events per gene. Let $\mathbf{\Lambda} = (\Lambda_1, \ldots, \Lambda_K)$. Then, recalling the conditional independence assumption between Markov chains, the joint distribution over all gene trees τ_k becomes

$$q(\tau_1 = v_1, \ldots, \tau_K = v_K | \Upsilon = u, \mathbf{\Lambda}) = \prod_{k=1}^{K} (\mathbf{P})_{uv_k}. \quad (8)$$

Calculating the probabilities in (8) requires numerical methods to determine the matrix exponential involving $\mathbf{P}_{\text{SPR}}$. These methods involve calculating the complete set of eigenvalues and eigenvectors of $\mathbf{P}_{\text{SPR}}$, requiring $\mathcal{O}(M^3)$ operations. Such procedures become quickly computationally prohibitive as N, and hence M, increases. As a consequence, numerical approximations may be necessary to develop weighted graph extensions to $\mathcal{G}_{\text{SPR}}$ directly. The weights in these extended graphs would be functions of unknown parameters and sampling these parameters would necessitate repetitive diagonalization.

Random walks with analytic solutions: An alternative to this computational barrier involves using random walks on graphs for which analytic solutions are known for any size M. To help find such solutions, Equation 7 demonstrates the close connection between a DTMC with a Poisson-distributed number of events and a CTMC. In fact, any such DTMC can be expressed as a unique CTMC, called the "continuized" version (D. Aldous and J. Fill, unpublished results). Analytic solutions for several weighted and unweighted CTMC processes on a complete graph are commonly used in phylogenetics. In a complete graph, all vertices are adjacent to all others. The most notable examples are the CTMC models for nucleotide substitution. The simplest model by Jukes and Cantor (1969) is unweighted. In the appendix, I present the multistep transition probability matrix $\mathbf{P}_{\text{GJC}}$ for a generalized Jukes-Cantor (GJC) model involving an arbitrary number of vertices M. Proposed by Kimura (1980), the next most sophisticated model for a complete graph is weighted. This model presupposes that the vertices are divided into two disjoint sets, $\boldsymbol{\mathcal{V}}_1 \cup \boldsymbol{\mathcal{V}}_2 = \boldsymbol{\mathcal{V}}$, and that transitions within and between $\boldsymbol{\mathcal{V}}_1$ and $\boldsymbol{\mathcal{V}}_2$ occur at varying rates. In terms of HGT, such a weighted random walk may prove useful to model varying rates of HGT between different groups of taxa. Letting $M_1 = |\boldsymbol{\mathcal{V}}_1|$, $M_2 = |\boldsymbol{\mathcal{V}}_2|$, and R equal the ratio of within- to between-transition rates, I present the multistep transition probability matrix $\mathbf{P}_{\text{GK}}$ given M_1, M_2, and R for a generalized Kimura (GK) model in the appendix.

Modeling differences across gene classes: I incorporate potential differences across genes in the expected number of HGT events Λ_k by employing a generalized linear model (GLM) approach (McCullagh and Nelder 1983). GLMs link the mean response, in this case Λ_k, to a set of linear predictors. First, I divide all K genes into one of C possible classes, where the definition of the classes depends on the specific research question at hand. To identify gene-class membership in the GLM, I construct a $K \times C$ design matrix $\mathbf{D} = (D_{kc})$, where matrix elements $D_{k1} = 1$ for all k, representing the baseline multiplier for the reference class, and

$$D_{kc} = \begin{cases} 1 & \text{if gene } k \in \text{class } c \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

for $c = 2, \ldots, C$, representing the offset multipliers for the remaining classes. Such a design matrix is standard in regression problems involving categorical dependent variables. I model

$$\log \Lambda_k = \sum_{c=1}^{C} \lambda_c D_{kc}, \quad (10)$$

where linear combinations of predictors $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_C)$ specify, on the log-scale, the expected number of HGT events for all classes. I complete the hierarchical prior specification by assuming

$$\boldsymbol{\lambda} \sim \text{Normal}(\mathbf{L}, \mathbf{\Psi}). \quad (11)$$

I set $\mathbf{L} = (-2, 0, \ldots, 0)$ and $\mathbf{\Psi} = \text{diag}(10, \ldots, 10)$. This provides a quite diffuse prior on $\boldsymbol{\lambda}$, with the median expected number of HGT events per gene ≈ 0.14 (Garcia-Vallve *et al.* 2000) for all classes.

As an example of how this GLM construction functions, consider the $C = 2$ classes case. Then,

$$\Lambda_k = \begin{cases} \exp(\lambda_1) & \text{if gene } k \in \text{class 1} \\ \exp(\lambda_1) \times \exp(\lambda_2) & \text{if gene } k \in \text{class 2.} \end{cases} \quad (12)$$

When $\lambda_2 = 0$, no difference across classes exists. Likewise, when $\lambda_2 < 0$, the expected number of HGT events per gene is smaller in class 2 than in class 1, and when $\lambda_2 > 0$, the expected number is larger.

STATISTICAL FRAMEWORK

Comparing the relative appropriateness of the various stochastic models for HGT proposed in preceding sections and testing for significant differences in the expected number of HGT events across genes can be accomplished using Bayesian model selection via Bayes factors. Bayes factors are the Bayesian analog of the likelihood-ratio test (LRT), but suffer from fewer difficulties than LRTs in discrete spaces, when comparing non-nested models and with sparse data (Suchard *et al.* 2001). A Bayes factor B_{10} measures the relative change in the support of the data $\mathbf{Y}$ in favor of one statistical model M_1 over another model M_0 and equals the ratio of the marginal likelihood $m(\mathbf{Y}|M_1)$ of M_1 over the marginal likelihood $m(\mathbf{Y}|M_0)$ of M_0 (Kass and Raftery 1995). To calculate Bayes factors, frequently more efficient methods than estimating the multidimensional

integrals hidden in the marginal likelihoods directly are available.

When models are nested, a relatively simple Bayes factor calculation is available via the Savage-Dickey ratio (Verdinelli and Wasserman 1995) and involves generating a posterior sample from the larger model only (Suchard *et al.* 2003b). For example, to assess the significance of differences across gene classes in the expected number of HGT events, let M_1 represent the unrestricted model proposed above. Nested within M_1 exists M_0, the equal-rates model, where $\lambda_c = 0$ for $c = 2, \ldots, C$. Further, the GJC model is nested within the GK model, as both are equal when $R = 1$.

On the other hand, the GJC and SPR models are nonnested, but both possess zero free parameters in their respective **P** matrices. For two arbitrary models M_0 and M_1 in situations like this, it is possible to estimate the posterior probabilities $p(M_0|\mathbf{Y})$ and $p(M_1|\mathbf{Y})$ by constructing a mixture model over the joint space of M_0 and M_1. By applying the Bayes theorem,

$$B_{10} = \frac{p(M_1|\mathbf{Y})}{p(M_0|\mathbf{Y})} \Big/ \frac{q(M_1)}{q(M_0)} = \frac{\text{Posterior odds}}{\text{Prior odds}}, \quad (13)$$

where $q(M_0)$ and $q(M_1)$ are the prior probabilities of models M_0 and M_1 in the mixture. Generally, I assume equal prior probabilities, $q(M_0) = q(M_1) = \frac{1}{2}$, when reporting posterior estimates. However, improved efficiency in estimating B_{10} can be garnered by adjusting these prior probabilities such that $p(M_0|\mathbf{Y}) \approx p(M_1|\mathbf{Y})$ (Carlin and Chib 1995; Suchard *et al.* 2002).

Models SPR and GK neither are nested nor contain the same number of free parameters. One might entertain constructing a reversible-jump Markov chain Monte Carlo (MCMC) sampler (Green 1995) over the joint space of these models to compute the Bayes factor in support of SPR over GK. However, a simpler algebraic solution exists given the two preceding Bayes factor calculations,

$$B_{\text{SPR,GK}} = \frac{B_{\text{SPR,GJC}}}{B_{\text{GJC,GK}}}. \quad (14)$$

To estimate all model parameters and Bayes factors, I employ MCMC. I further develop this MCMC algorithm and discuss its performance in the appendix.

EXAMPLE

To illustrate these stochastic models for HGT and methods to test hypotheses about them, I examine a large set of orthologous, prokaryotic genes collected by Jain *et al.* (1999). The data consist of $K = 144$ separate gene alignments. Each alignment contains orthologous copies of a single gene from six prokaryotes. These prokaryotes are: *Aquifex aeolicus* (Aa), an early branching thermophilic eubacterium; *Escherichia coli* (Ec), a proteobacterium; *Synechocystis 6803* (S6), a cyanobacterium; *Bacillus subtilis* (Bs), a gram-positive bacterium; *Methanococcus jannaschii* (Mj), a methanogen; and *Archaeoglobus fulgidus* (Af), a thermophilic sulfate-reducing methanogen relative. The first four organisms are Eubacteria, while the last two are Archaea. Jain *et al.* (1999) construct the gene alignments on the basis of amino acid translations, assuming a star tree to reduce alignment bias (Lake 1991), and classify each gene into one of two distinct classes, informational and operational genes (Rivera *et al.* 1998). Table 1 lists the functional characteristics of the genes that fall into each class. As a generalization, informational gene products interact in large complex systems; this is especially true of the translational and transcriptional apparatuses. On the other hand, most operational gene products function independently or in small protein assemblies. In total, Jain *et al.* (1999) assign 56 genes as informational and 88 as operational, employ these genes to examine the complexity hypothesis, and find support for higher levels of HGT among the operational genes as compared to the informational genes.

TABLE 1

Functional definitions of two distinct gene classes, adopted from Rivera *et al.* (1998)

Gene-class c =	
1. Informational	2. Operational
Transcription	Regulatory genes
Translation	Cell envelope proteins
tRNA synthetases	Intermediary metabolism
GTPases/vacuolar ATPase homologs	Biosynthesis of amino acids, fatty acids phospholipids, cofactors, and nucleotides

I parallel the above analysis by assuming that the number of the different gene classes $C = 2$. I let class $c = 1$ represent the informational genes and class $c = 2$ represent the operational genes. To further maintain consistency with Jain *et al.* (1999), I exclude third codon position nucleotides from all alignments and assume that first and second codon position nucleotides are evolving independently under the same process for each gene.

Selection of stochastic model: I begin by comparing the relative likelihoods of the three different stochastic models, SPR, GJC, and GK. For the GK model, I define my two disjoint sets of trees as (1) those that support a split between the four Eubacteria and the two Archaea, $\boldsymbol{\nu}_1$, and (2) those that do not, $\boldsymbol{\nu}_2$. These definitions offer a first approximation to modeling differing rates of HGT within life domains and across domains in this example. HGT events that start and end in set $\boldsymbol{\nu}_1$ are within domain transfers, while events that start in $\boldsymbol{\nu}_1$ and end in $\boldsymbol{\nu}_2$, or vice versa, are across domain transfers.

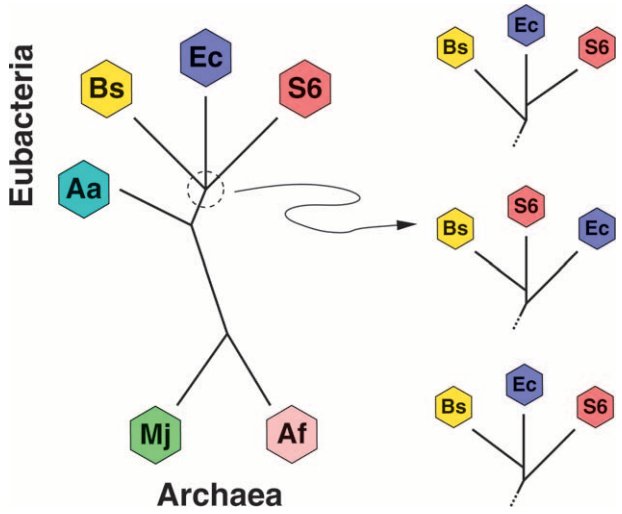

FIGURE 3.—Species tree relating six prokaryotes. Species are: *Aquifex aeolicus* (Aa), *Escherichia coli* (Ec), *Synechocystis 6803* (S6), *Bacillus subtilis* (Bs), *Methanococcus jannaschii* (Mj), and *Archaeoglobus fulgidus* (Af). Branch order of three Eubacteria Ec, S6, and Bs is under debate, leading to three possible subtrees (shown on right).

The $\log_{10}$ Bayes factor in favor of SPR over GJC and the $\log_{10}$ Bayes factor in favor of GK over GJC are

$$\log_{10} B_{\text{SPR,GJC}} = 19.2 \quad \text{and} \quad \log_{10} B_{\text{GK,GJC}} = 5.7. \tag{15}$$

Figure 4a illustrates the scaled regeneration quantile (SRQ) plot for estimating the relative posterior probabilities used to calculate $\log_{10} B_{\text{SPR,GJC}}$. No substantial deviation from the slope = 1 line implies the MCMC chain is mixing sufficiently to generate this estimate. Combining the results in (15), I calculate the $\log_{10}$ Bayes factor in favor of SPR over GK as

$$\log_{10} B_{\text{SPR,GK}} = 19.2 - 5.7 = 13.5. \tag{16}$$

Considering these Bayes factor estimates, the data strongly reject (KASS and RAFTERY 1995) the two complete graph models with analytic solutions in favor of the more biologically plausible process based on the SPR operator. However, the GJC and GK models should not be discounted completely; their computational complexity does not increase with increasing number of taxa N and they can offer some insight into the underlying biological processes. For example, the Bayes factor in favor of GK over GJC offers some indirect support for differing HGT rates within domains rather than across domains. One caveat should be kept in mind to keep from drawing too strong a conclusion from this finding—the unbalanced study design with only two Archaea precludes identifying HGT events within that domain. All further results in this article are based on the SPR model.

Estimating the species tree: Figure 3 displays the currently accepted species tree relating the six prokaryotes studied here. The four Eubacteria and two Archaea form two distinct clades (FENG *et al.* 1997) and Aa is the earliest branching species of the Eubacteria studied (DECKERT *et al.* 1998). The branching order of the remaining three Eubacteria Ec, S6, and Bs is more ambiguous (GIOVANNONI *et al.* 1996). The three possible resolutions of this trifurcation are depicted on the right side of Figure 3. Much of the debate surrounding the trifurcation depends on data choice and reconstruction methodology. For example, the top resolution produces species tree $\Upsilon_{\text{Ec-S6}}$ that places Ec and S6 as nearest neighbors. Protein synthesis elongation factor (EF) Tu gene reconstructions support this tree (LAKE and RIVERA 1996) and JAIN *et al.* (1999) fix $\Upsilon_{\text{Ec-S6}}$ as their reference tree in their analysis. Reconstructions of 16S rRNA phylogeny support the middle resolution of species tree $\Upsilon_{\text{Bs-S6}}$ (COLE *et al.* 2003) with Bs and S6 as nearest neighbors. The final resolution of species tree $\Upsilon_{\text{Ec-Bs}}$ gains support from reconstructions of phenylalanyl-tRNA synthetase (TEICHMANN and MITCHISON 1999). However, even these three critical genes are subject to HGT (WOLF *et al.* 1999; ZAP *et al.* 1999; KE *et al.* 2000) and their reconstructed phylogenies may inaccurately represent the true species tree.

On the basis of the SPR model for HGT, I infer $\Upsilon_{\text{Bs-S6}}$ as the most likely species tree with >0.999 posterior probability. The two other resolutions, $\Upsilon_{\text{Ec-Bs}}$ and $\Upsilon_{\text{Ec-S6}}$, are the second and third most likely species trees, respectively. To estimate the Bayes factors in favor of $\Upsilon_{\text{Bs-S6}}$ against $\Upsilon_{\text{Ec-Bs}}$ and $\Upsilon_{\text{Ec-S6}}$, I judiciously reweight my prior probabilities on trees $\mathbf{z}$ and calculate

$$\log_{10} B_{\text{Bs-S6,Ec-Bs}} = 7.3 \quad \text{and} \quad \log_{10} B_{\text{Ec-Bs,Ec-S6}} = 5.9. \tag{17}$$

Similar to the back calculation completed in previous section, I estimate

$$\log_{10} B_{\text{Bs-S6,Ec-S6}} = 7.3 + 5.9 = 13.2, \tag{18}$$

while direct calculation of $\log_{10} B_{\text{Bs-S6,Ec-S6}}$ using the sampler yields approximately the same result. Figure 4, b–d, depicts the SRQ plots relevant to these Bayes factor calculations. Again, the MCMC chain appears well mixing. Although the posterior support for $\Upsilon_{\text{Ec-Bs}}$ and $\Upsilon_{\text{Ec-S6}}$ initially appears quite small, on a relative scale it is not; probabilities for the remaining 102 trees are >15 orders of magnitude smaller.

Data sets as large as the $K = 144$ gene alignments from JAIN *et al.* (1999) are currently rare. Consequentially, I examine via simulation the number of alignments necessary to identify the species tree under the SPR model. Under this simulation, I randomly sample without replacement a fixed number of gene alignments K and then estimate the posterior support for $\Upsilon_{\text{Bs-S6}}$, assuming this is the true species-tree. I repeat this simulation 20 times for each value of K. For $K = 2$, the expected posterior probability of $\Upsilon_{\text{Bs-S6}} = 0.14$. This estimate is approaching its prior value, signifying appropriate MCMC sampling with limited data. Approximately $K = 50$ gene alignments are required to achieve an expected

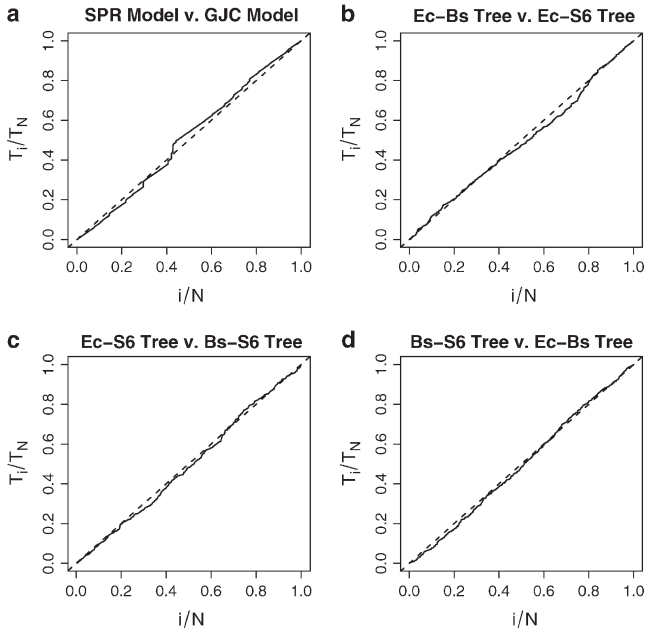

FIGURE 4.—Scaled regeneration quantile (SRQ) plots to assess MCMC sampler performance when estimating four relative posterior probabilities. Plot a was generated when comparing the SPR and GJC stochastic models. Plots b–d were generated when comparing the three most probable species trees. No substantial deviation in the slopes from 1 (dashed lines) implies that the chains are mixing well enough to produce stable estimates.

posterior probability ≥ 0.80 and $K = 70$ are required for ≥ 0.90.

Hierarchical estimates of evolutionary pressures: Table 2 presents the posterior estimates of the across-gene-level parameters used to pool information about (α_k, γ_k, μ_k, $\boldsymbol{\pi}_k$). The table also lists posterior estimates of

$$A' = \exp(A + \tfrac{1}{2}\sigma_A^2),$$
$$G' = \exp(G + \tfrac{1}{2}\sigma_G^2),$$
$$M' = \exp(M + \tfrac{1}{2}\sigma_M^2). \quad (19)$$

These transformed variables report the across-gene-level averages of the two transition:transversion ratios and expected divergence on their usual, instead of log, scale. As seen from Table 2, the average transition:transversion ratio for purines A' is significantly different from the ratio for pyrimidines G', as the ratios' 95% Bayesian credible intervals (BCIs) do not overlap, and both ratios are greater than one. This supports the use of the TN93 model for nucleotide substitution over a more restricted model. Estimates of A, G, and $\boldsymbol{\Pi}$ are consistent with a previous study using a subset of the data in a hierarchical framework (SUCHARD *et al.* 2003a). Also in comparison to this previous study, differences in estimates of M, σ_A^2, σ_G^2, σ_M^2, and N_Π all trend in the correct directions given the increase in the number of taxa and genes fit here.

Varying rates of HGT across gene classes: Figure 5 displays model estimates for the linear predictors λ_1 and λ_2 and for the expected number of HGT events per gene, Λ_k, for the informational and operational gene classes. The two top plots display histograms of the posterior samples of λ_1 (left) and λ_2 (right). These plots also include normal approximations to the posterior (solid lines) and prior densities (dashed lines). Examining the plot on the right, the prior density at $\lambda_2 = 0$ (dotted vertical line) is considerably higher than the normal approximation to the posterior density. Further, the 95% BCI of $\lambda_2 = (0.27–1.15)$ and does not cover zero. Both observations support the hypothesis that $\lambda_2 \neq 0$ and, hence, that rates of HGT differ between informational and operational genes. Formally, the Bayes factor in favor of differing rates is given by the Savage-Dickey ratio. The $\log_{10}$ Bayes factor,

$$\log_{10} B_{\neq \text{rates}, = \text{rates}} = 0.9, \quad (20)$$

offers substantial support (KASS and RAFTERY 1995) in favor of differing rates.

The bottom plot in Figure 5 transforms λ_1 and λ_2 into the expected number of HGT events per gene and displays histograms of the posterior samples of these quantities. Depicted in dark shading is Λ_k for the operational genes and depicted in light shading is Λ_k for the informational genes. Although Λ_k for operational genes is significantly greater than Λ_k for informational genes from the argument above, a small amount of overlap is

TABLE 2

Hierarchical across-gene-level estimates of evolutionary pressures

Log-scale central tendencies			Natural-scale central tendencies			Measures of precision		
Parameter	Mean	(95% BCI)	Parameter	Mean	(95% BCI)	Parameter	Mean	(95% BCI)
A	0.48	(0.4–0.52)	A'	1.65	(1.59–1.71)	$1/\sigma_A^2$	26.59	(20.29–33.77)
G	0.23	(0.180–0.28)	G'	1.29	(1.23–1.35)	$1/\sigma_G^2$	20.28	(14.67–27.15)
M	−1.67	(−1.74–−1.60)	M'	0.19	(0.18–0.21)	$1/\sigma_M^2$	17.02	(11.65–23.93)
			Π_A	0.35	(0.35–0.35)	N_Π	542.66	(455.47–639.81)
			Π_G	0.30	(0.29–0.30)			
			Π_C	0.17	(0.16–0.17)			
			Π_T	0.19	(0.18–0.19)			

Posterior means and 95% Bayesian credible intervals (BCIs) are reported for each parameter.

observed (solid shading) between these marginal histograms. This overlap results from the high negative correlation between λ_1 and λ_2 (data not shown) and illustrates the need for caution in making inference on the basis of marginal posterior summaries alone.

REMARKS

In this article, I proposed a simple class of stochastic models for HGT. The models are based on a random walk process in tree space and allow for the development of a joint distribution over multiple gene trees given an unknown species tree. I consider two general forms of random walks. The first stems from subtree transfer operations, in particular the SPR operator that mirrors the observed effects that HGT has on an inferred tree. The second form is based on walks over complete graphs and offers numerically tractable solutions for increasing number of taxa. I fit these models using a Bayesian framework to data from six prokaryotes. I find strongest support for the species tree that places Bs and S6 as nearest neighbors. This tree is supported by 16S rRNA reconstructions, but differs from the EF-Tu tree assumed by Jain *et al.* (1999). I demonstrate the flexibility of these stochastic models to test competing ideas about HGT by examining the complexity hypothesis and find support for increased HGT of operational genes compared to informational genes. This latter finding remains unchanged if I fix the species tree to equal the EF-Tu tree (data not shown).

The specific stochastic models for HGT developed in this article have important limitations. First and foremost, the random walks explore only the discrete, topological portion of tree space and do not consider changes in branch lengths between trees as part of the underlying HGT process. As a result, HGT between nearest neighbors in a tree remains unidentified as this process does not result in a change in the topological configuration of the tree. Model extensions that consider a continuous random drift process on the joint space of $(\tau, \mathbf{t})$ (Billera *et al.* 2001) may circumvent this shortfall. For a related problem involving coalescence, Yang (2002) shows that including branch lengths $\mathbf{t}$ into the probabilistic model across loci improves power. Additionally, I assume that the K DTMCs representing the random walks of the gene trees τ_k away from the species tree Υ are conditionally independent given Υ. This assumption implies that the evolutionary histories of all genes are unlinked, while evidence for the HGT of, at a minimum, complete operons abounds in prokaryotes (Koonin *et al.* 2001). Possible modeling aspects include allowing for linked or partially linked genes.

HGT is not the only process that may cause incongruence between gene trees. Although the effects of lineage sorting should be minor given the extensive divergence between the species studied here, the inclusion of paralogous genes copies within the orthologous alignments may mislead inference. Also important, stochastic error due to sparse phylogenetic data, evolutionary model misspecification, and parallel/convergent evolution can falsely produce incongruence between trees (Cao *et al.* 1998). These effects should upwardly bias the inferred number of HGT events. However, I suspect this bias is less than one HGT event per gene as only a modest percentage of genes should be affected and the error should produce just minor changes in the inferred tree. There is no *a priori* reason to suspect that this bias differs between the informational and operational gene classes; so the bias does not affect the relative difference between classes in HGT rates and inference regarding the complexity hypothesis.

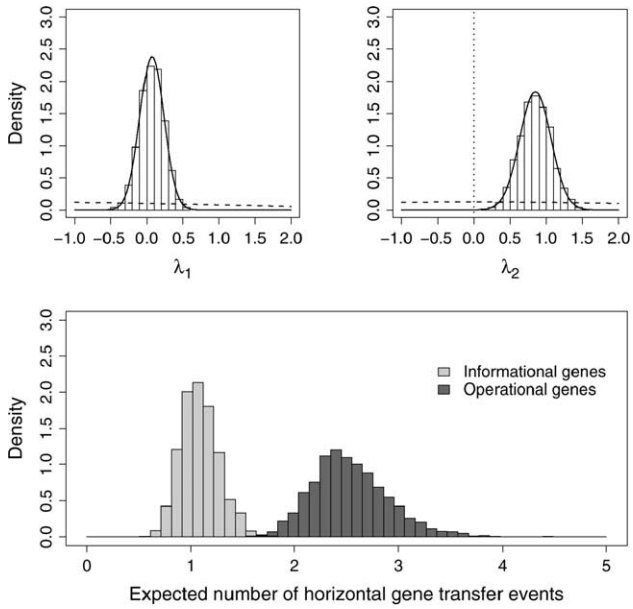

Figure 5.—Analysis of the complexity hypothesis. The top two plots depict the posterior distributions of linear predictors λ_1 and λ_2 using histograms and normal approximations (solid lines). Also shown are the predictors' prior densities (dashed lines). Greater prior than posterior density at $\lambda_2 = 0$ (dotted line) supports a difference in HGT rates between gene classes. The bottom plot depicts the posterior distributions of the expected number of HGT events per gene for informational genes (light shading) and operational genes (dark shading).

For the SPR model, numerical approximations to the matrix exponentials involving the multistep transition probability matrix $\mathbf{P}_{\text{SPR}}$ may offer promise in handling research problems with larger numbers of taxa N (Moler and Van Loan 2003). As N increases, the square dimensions of $\mathbf{P}_{\text{SPR}}$ grow superexponential, while the size of the neighborhood of each vertex grows only as $\mathcal{O}(N^2)$. As a consequence, $\mathbf{P}_{\text{SPR}}$ becomes increasingly sparse. In this situation, the number of unique eigenvalues increases substantially slower than the matrix's dimension. Krylov subspace techniques (Sidje and Stewart 1999) may stretch computational limits upward to $N = 8$ or more.

In spite of these limitations, these stochastic models for HGT offer several advantages over previous approaches to studying HGT using multiple orthologous

gene alignments. Under these stochastic models, the species tree is an unknown parameter that may be either integrated out of the analysis as a nuisance parameter or estimated jointly with the multiple gene trees. Joint analysis decreases the possibility of bias introduced through fixing the species tree when knowledge about it is uncertain. A stochastic approach also overcomes the bias inherent in parsimony-like estimation. Further, the hierarchical framework in which the stochastic model sits enables the borrowing of strength in the estimation of all gene-partition-level estimates including the gene trees themselves. Finally, and most importantly, stochastic models lend themselves well to formal statistical testing, with no need for *ad hoc* procedures. The ability to compare differing models for HGT will continue to shed further insight into the underlying biological process.

I thank the Lake lab, in particular Jon Moore and Jim Lake, for stimulating my interest in HGT, for many provocative discussions, and for providing the SPR adjacency matrices and prokaryote sequences used in this study. I also thank John Huelsenbeck for his insights into HGT and Janet Sinsheimer and Vladimir Minin for commenting on this manuscript. Fred Fox and the National Science Foundation grant 9987641-sponsored University of California, Los Angeles, Training Program in Bioinformatics graciously made possible the computing facilities to fit all 144 alignments simultaneously. The complete data set is available to interested readers at http://www.biomath.medsch.ucla.edu/msuchard/datasets.html. I am supported in part by National Institutes of Health grants GM08042 and GM068955 and U.S. Public Health Service grant CA16042.

LITERATURE CITED

Allen, B., and M. Steel, 2001 Subtree transfer operations and their induced matrices on evolutionary trees. Ann. Combinatorics **5:** 1–15.

Bergsmedh, A., A. Szeles, M. Henriksson, A. Bratt, M. Folkman *et al.*, 2001 Horizontal transfer of oncogenes by uptake of apoptotic bodies. Proc. Natl. Acad. Sci. USA **98:** 6407–6411.

Billera, L., S. Holmes and K. Vogtmann, 2001 Geometry of the space of phylogenetic trees. Adv. Appl. Math. **27:** 733–767.

Brown, J., 2003 Ancient horizontal gene transfer. Nat. Rev. Genet. **4:** 121–132.

Cao, Y., A. Janke, P. Waddell, M. Westerman, O. Takenaka *et al.*, 1998 Conflict among individual mitochondrial proteins in resolving the phylogeny of Eutherian orders. J. Mol. Evol. **47:** 307–322.

Carlin, B., and S. Chib, 1995 Bayesian model choice via Markov chain Monte Carlo methods. J. R. Stat. Soc. Ser. B **57:** 473–484.

Cole, J., B. Chai, T. Marsh, R. Farris, Q. Wang *et al.*, 2003 The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res. **31:** 442–443.

Deckert, G., P. Warren, T. Gaasterland, W. Young, A. Lenox *et al.*, 1998 The complete genome of the hyperthermophilic bacterium *aquifex aeoclicus.* Nature **392:** 353–358.

Doolittle, W., 1999 Lateral gene transfer, genome surveys and the phylogeny of prokaryotes. Science **286:** 1443a.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Feng, D., G. Cho and R. Doolittle, 1997 Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94:** 13028–13033.

Garcia-Vallve, S., A. Romeu and J. Palau, 2000 Horizontal gene transfer in bacterial and archeal complete genomes. Genome Res. **10:** 1719–1725.

Gelman, A., G. Roberts and W. Gilks, 1996 Efficient Metropolis jumping rules, pp. 599–608 in *Bayesian Statistics*, Vol. 5, edited by J. Bernardo, J. Berger, A. Dawid and A. Smith. Oxford University Press, Oxford.

Giovannoni, S., M. Rapp, D. Gordon, E. Urbach, M. Suzuki *et al.*, 1996 Ribosomal RNA and the evolultion of bacterial diversity, pp. 63–85 in *Evolution of Microbial Life*, edited by D. Roberts, P. Sharp, G. Alderson and M. Collins. Cambridge University Press, Cambridge, UK.

Green, P., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Hein, J., 1990 Reconstructing evolution of sequences subjects to recombination using parsimony. Math. Biosci. **98:** 185–200.

Hein, J., 1993 A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36:** 396–405.

Huelsenbeck, J., F. Ronquist, R. Nielsen and J. Bollback, 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294:** 2310–2314.

Jain, R., M. Rivera and J. Lake, 1999 Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA **96**: 3801–3806.

Jain, R., M. Rivera, J. Moore and J. Lake, 2002 Horizontal gene transfer in microbial genome evolution. Theor. Popul. Biol. **61:** 489–495.

Jukes, T., and C. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammaliam Protein Metabolism*, edited by H. Munro. Academic Press, New York.

Kass, R., and A. Raftery, 1995 Bayes factors. J. Am. Stat. Assoc. **90:** 773–795.

Ke, D., M. Boissinot, A. Huletsky, F. Picard, J. Frenette *et al.*, 2000 Evidence for horizontal gene transfer in evolution of elongation factor Tu in enterococci. J. Bacteriol. **182:** 6913–6920.

Kimura, M., 1980 A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

Koonin, E., K. Makarova and L. Aravind, 2001 Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. **55:** 709–742.

Koski, L., R. Morton and G. Golding, 2001 Codon bias and base composition are poor indicators of horizontally transferred genes. Mol. Biol. Evol. **18:** 404–412.

Lake, J., 1991 The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. **8:** 378–385.

Lake, J., and M. Rivera, 1996 The prokaryotic ancestry of eukaryotes, pp. 87–108 in *Evolution of Microbial Life*, edited by D. Roberts, P. Sharp, G. Alderson and M. Collins. Cambridge University Press, Cambridge, UK.

Lawrence, J., 1999 Gene transfer, speciation and the evolution of bacterial genomes. Curr. Opin. Microbiol. **2:** 519–523.

Lawrence, J., and H. Ochman, 1997 Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:** 383–397.

Leverstein-van Hall, M., A. Box, H. Blok, A. Pauuw, A. Fluit *et al.*, 2002 Evidence of extensive interspecies transfer of integron-mediated antimicrobial resistance genes among multidrug-resistant Enterobacteriaceae in a clinical setting. J. Infect. Dis. **186:** 49–56.

Li, S., D. Pearl and H. Doss, 2000 Phylogenetic tree construction using Markov chain Monte Carlo. J. Am. Stat. Assoc. **95:** 493–508.

Liu, J., 1994 The collasped Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J. Am. Stat. Assoc. **89:** 958–966.

Maddison, W., 1997 Gene trees in species trees. Syst. Biol. **46:** 523–536.

Mau, B., M. Newton and B. Larget, 1999 Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55:** 1–12.

McCullagh, P., and J. Nelder, 1983 *Generalized Linear Models: Monographs on Statistics and Applied Probability.* Chapman & Hall, New York..

Mirkin, B., T. Fenner, M. Galperin and E. Koonin, 2003 Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. **3:** 2.

Moler, C., and C. Van Loan, 2003 Nineteen dubious ways to com-

pute the exponential of a matrix, twenty-five years later. Soc. Ind. Appl. Math. Rev. **45:** 3–49.
Mykland, P., L. Tierney and B. Yu, 1995 Regeneration in Markov chain samplers. J. Am. Stat. Assoc. **90:** 233–241.
Page, R., 2000 Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol. Phylogenet. Evol. **14:** 89–106.
Ragan, M., 2001 Detection of lateral gene transfer among microbial genomes. Curr. Opin. Genet. Dev. **11:** 620–626.
Rivera, M., R. Jain, J. Moore and J. Lake, 1998 Genomic evidence of two functionally distinct gene classes. Proc. Natl. Acad. Sci. USA **95:** 6239–6244.
Roberts, G., and S. Sahu, 1997 Updating schemes, correlation structure, blocking and parameterization of the Gibbs sampler. J. R. Stat. Soc. Ser. B **59:** 291–317.
Robinson, D., 1971 Comparison of labeled trees with valency three. J. Comb. Theor. Ser. B **11:** 105–119.
Sidje, R., and W. Stewart, 1999 A numerical study of large sparse matrix exponentials arising in Markov chains. Comput. Stat. Data Anal. **29:** 345–368.
Sinsheimer, J., J. Lake and R. Little, 1996 Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. Biometrics **52:** 193–210.
Snel, B., P. Bork and M. Huynen, 1999 Genome phylogeny based on gene content. Nat. Genet. **21:** 108–110.
Suchard, M., R. Weiss and J. Sinsheimer, 2001 Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. **18:** 1001–1013.
Suchard, M., R. Weiss, K. Dorman and J. Sinsheimer, 2002 Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst. Biol. **51:** 715–728.
Suchard, M., C. Kitchen, J. Sinsheimer and R. Weiss, 2003a Hierarchical phylogeneic models for analyzing multipartite sequence data. Syst. Biol. **52:** 649–664.
Suchard, M., R. Weiss and J. Sinsheimer, 2003b Testing a molecular clock without an outgroup: derivations of induced priors on branch length restrictions in a Bayesian framework. Syst. Biol. **52:** 48–54.
Swofford, D., G. Olsen, P. Waddell and D. Hillis, 1996 Phylogenetic inferences, pp. 407–514 in *Molecular Systematics*, Ed. 2, edited by D. Hillis, C. Moritz and B. Mable. Sinauer Associates, Sunderland, MA.
Syvanen, M., 1994 Horizontal gene transfer: evidence and possible consequences. Annu. Rev. Genet. **28:** 237–261.
Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.
Taylor, J., A. Siqueira and R. Weiss, 1996 The cost of adding parameters to a model. J. R. Soc. Stat. Ser. B **58:** 593–607.
Teichmann, S., and G. Mitchison, 1999 Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol. **49:** 98–107.
Verdinelli, I., and L. Wasserman, 1995 Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J. Am. Stat. Assoc. **90:** 614–618.
Woese, C., 2000 Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. USA **97:** 8392–8396.
Wolf, Y., L. Aravind, N. Grishin and E. Koonin, 1999 Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. **9:** 689–710.
Yang, Z., 2002 Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics **162:** 1811–1823.
Yang, Z., and B. Rannala, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14:** 717–724.
Zap, W., Z. Zhang and Y. Wang, 1999 Distinct types of rRNA operons exist in the genome of the actinomycete *thermomonospora chromogena* and evidence for horizontal gene transfer of an entire rRNA operon. J. Bacteriol. **181:** 5201–5209.

Communicating editor: J. Hein

APPENDIX

Complete models: To determine the multistep transition probability matrix $\mathbf{P}_{GJC}$ for the GJC model with $M \geq 2$ states, I first recall that

$$\mathbf{P}_{GJC} = \exp(\Lambda_k \mathbf{Q}_{GJC}) \tag{A1}$$

is generated from an unweighted complete graph. As a complete graph, it is trivially connected and, therefore, has a unique stationary distribution. This distribution is $(1/M, \ldots, 1/M)$.

To determine the eigenvalues of $\mathbf{Q}_{GJC}$, I write

$$\mathbf{Q}_{GJC} = \frac{1}{M-1}\mathbf{J} - \frac{M}{M-1}\mathbf{I}, \tag{A2}$$

where $\mathbf{Q}_{GJC}$ is scaled such that Λ_k is expressed in terms of the expected number of HGT events per gene, $\mathbf{J}$ is the $M \times M$ matrix of all ones, and $\mathbf{I}$ is the $M \times M$ identity matrix. Matrix $\mathbf{J}$ has a rank of one and, therefore, one nonzero eigenvalue that equals $M/(M-1)$. Given the eigenvalues of $\mathbf{J}$ and expression (A2), the M eigenvalues of $\mathbf{Q}_{GJC}$ become

$$\left(0, \frac{-M}{M-1}, \ldots, \frac{-M}{M-1}\right). \tag{A3}$$

Like the standard Jukes-Cantor model, where $M = 4$, the GJC model for any $M \geq 2$ continues to have only two distinct eigenvalues. Conceptually this results because the qualitative behavior of the underlying Markov chain does not change as the size of the state-space increases.

By letting $\Lambda_k \rightarrow \infty$, I see that the stationary distribution is the eigenvector corresponding to the 0 eigenvalue. By examining the other limiting case where $\Lambda_k = 0$ and considering the initial conditions, algebraic rearrangement yields

$$(\mathbf{P}_{\mathrm{GJC}})_{uv} = \begin{cases} \frac{1}{M} + \frac{M-1}{M}\exp(-\frac{M}{M-1}\Lambda_k) & \text{if } u = v \\ \frac{1}{M} - \frac{1}{M}\exp(-\frac{M}{M-1}\Lambda_k) & \text{otherwise.} \end{cases} \tag{A4}$$

The state-space of the GK model is partitioned into two disjoint sets $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$. Let $M_1 = |\boldsymbol{\nu}_1|$ and $M_2 = |\boldsymbol{\nu}_2|$, where $M_1 + M_2 = M$, and let R be the ratio of rates for transitions within a structural set to transitions between sets. Then, following arguments similar to those above, one can find the multistep transition probability matrix $\mathbf{P}_{\mathrm{GK}}$ for the GK model.

If $u \in \boldsymbol{\nu}_1$, then

$$(\mathbf{P}_{\mathrm{GK}})_{uv} = \begin{cases} \frac{1}{M} + \frac{M_1-1}{M_1}\exp(-\phi_1\gamma\Lambda_k) + \frac{M_2}{M_1M}\exp(-\phi_2\gamma\Lambda_k) & \text{if } u = v \\ \frac{1}{M} - \frac{1}{M_1}\exp(-\phi_1\gamma\Lambda_k) + \frac{M_2}{M_1M}\exp(-\phi_2\gamma\Lambda_k) & \text{else if } v \in \boldsymbol{\nu}_1 \\ \frac{1}{M} - \frac{1}{M}\exp(-\phi_2\gamma\Lambda_k) & \text{otherwise,} \end{cases} \tag{A5}$$

where

$$\begin{aligned} \gamma &= \frac{M}{[M_1(M_1-1) + M_2(M_2-1)]R + 2M_1M_2}, \\ \phi_1 &= M_1R + M_2, \\ \phi_2 &= M. \end{aligned} \tag{A6}$$

By symmetry, if $u \in \boldsymbol{\nu}_2$, then

$$(\mathbf{P}_{\mathrm{GK}})_{uv} = \begin{cases} \frac{1}{M} + \frac{M_2-1}{M_2}\exp(-\phi_3\gamma\Lambda_k) + \frac{M_1}{M_2M}\exp(-\phi_2\gamma\Lambda_k) & \text{if } u = v \\ \frac{1}{M} - \frac{1}{M_2}\exp(-\phi_3\gamma\Lambda_k) + \frac{M_1}{M_2M}\exp(-\phi_2\gamma\Lambda_k) & \text{else if } v \in \boldsymbol{\nu}_2 \\ \frac{1}{M} - \frac{1}{M}\exp(-\phi_2\gamma\Lambda_k) & \text{otherwise,} \end{cases} \tag{A7}$$

where $\phi_3 = M_2R + M_1$. For $R \neq 1$, note that there are four unique eigenvalues when $M_1 \neq M_2$ and three unique eigenvalues otherwise. This is consistent with the standard Kimura model, in which $M_1 = M_2 = 2$ with three unique eigenvalues.

Sampling algorithm: For each gene-partition k, let $\boldsymbol{\theta}_k = (\tau_k, \mathbf{t}_k, \alpha_k, \gamma_k, \mu_k, \boldsymbol{\pi}_k)$ and, then, assemble $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ to be the collection of all gene-level parameters. To specify the hierarchical prior parameters, let $\phi = (\mathbf{V}, \boldsymbol{\Sigma}, \boldsymbol{\Pi}, N_\Pi, \Upsilon, \boldsymbol{\lambda})$. Across-gene-level parameters ϕ also include R when considering the GK model and mixing parameter $\psi \in \{0, 1\}$ when comparing models SPR and GJC. I employ a MCMC approach to sample from each model's joint posterior distribution, $p(\boldsymbol{\theta}, \phi|\mathbf{Y})$. I generate samples from these posteriors using two nested Metropolis-within-Gibbs cycles, as laid out in Suchard *et al.* (2003a) for hierarchical phylogenetic models. The outer cycle first iterates over gene partitions k and then over the parameters in ϕ. Within each gene partition k, the inner cycle proceeds over the parameters in $\boldsymbol{\theta}_k$. With the exception of proposals for Υ, $\boldsymbol{\lambda}$, R, and ψ, all parameter proposals follow those in Suchard *et al.* (2003a).

The multinomial prior placed on Υ is conjugate to its sampling density. As a result, it is possible to Gibbs sample Υ from its full conditional distribution for moderately small M. This full conditional distribution remains multinomial with M state probabilities given by

$$p(\Upsilon = u|\mathbf{Y}, \boldsymbol{\Omega}_{-\Upsilon}) = \frac{\prod_{k=1}^{K}(\mathbf{P})_{uv_k}z_u}{\sum_{w\in\mathfrak{v}}\prod_{k=1}^{K}(\mathbf{P})_{wv_k}z_w}, \tag{A8}$$

where $\tau_k = v_k$ for all k and $\boldsymbol{\Omega}_{-\Upsilon}$ is the vector of all model parameters $(\boldsymbol{\theta}, \phi)$ excluding Υ. Similar to the reweighted prior approach to estimate ψ, varying $\mathbf{z}$ can improve sampling efficiency when estimating the relative posterior probabilities of specific species trees Υ.

I draw the transition ratio R and linear predictors $\boldsymbol{\lambda}$ via separate Metropolis-Hastings proposals. For R, I propose

new parameter values by generating a normal random variate centered at the current value of R with a tunable variance s_R^2. Given the high degree of correlation between column vectors in the design matrix $\mathbf{D}$, I expect the posterior distribution of $\boldsymbol{\lambda}$ to also exhibit strong correlation. This expectation stems from a normal linear regression approximation to $p(\exp(\boldsymbol{\lambda})|\boldsymbol{\Lambda})$ that has a variance-covariance structure proportional to $(\mathbf{D}'\mathbf{D})^{-1}$. As a consequence, component-by-component updating of λ_c in $\boldsymbol{\lambda}$ should lead to a slowly mixing MCMC chain (Roberts and Sahu 1997). To help ensure adequate mixing, I propose all λ_c simultaneously using a multivariate normal random variate centered at the current value of $\boldsymbol{\lambda}$ with a tunable variance-covariance matrix $\mathrm{diag}(s_{\lambda_1}^2, \ldots, s_{\lambda_C}^2)\boldsymbol{\Xi}$. I adjust the tunable variances such that proposals have acceptance rates of 30–40% (Gelman *et al.* 1996) and fix the correlation matrix $\boldsymbol{\Xi}$ approximately equal to the posterior correlation of $\boldsymbol{\lambda}$ determined by a trial chain.

When comparing HGT models using a mixture approach, I sample the mixing parameter ψ directly from its full conditional distribution in a Gibbs step,

$$\psi|\mathbf{Y}, \boldsymbol{\Omega}_{-\psi} \sim \mathrm{Bernoulli}(a), \tag{A9}$$

where

$$a = \frac{b_1 q(M_1)}{b_0 q(M_0) + b_1 q(M_1)}$$
$$b_i = \prod_{k=1}^{K} (\boldsymbol{P}_{M_i})_{uv_k}, \tag{A10}$$

for $i = 0, 1$, $\Upsilon = u$, and $\tau_k = v_k$. Values a may be saved at each iteration and used to construct a Rao-Blackwellized estimator for $p(M_1|\mathbf{Y})$ (Suchard *et al.* 2003a).

Finally, the inferred number of HGT events E_k for the SPR model can be recovered after posterior simulation. The full conditional distribution

$$p(E_k|\mathbf{Y}, \boldsymbol{\Omega}_{-E_k}) = p(E_k|\tau_k, \Upsilon, \Lambda_k),$$
$$= \frac{(\mathbf{A}^{E_k})_{uv_k} e^{-\Lambda_k}(\Lambda_k^{E_k}/E_k!)}{(\mathbf{P})_{uv_k}}, \tag{A11}$$

where $\Upsilon = u$ and $\tau_k = v_k$. Since $(\mathbf{A}^{E_k})_{uv_k} \leq 1$,

$$p(E_k|\mathbf{Y}, \boldsymbol{\Omega}_{-E_k}) \leq \frac{1}{(\mathbf{P})_{uv_k}} p(E^*|\Lambda_k), \tag{A12}$$

where

$$E^* \sim \mathrm{Poisson}(\Lambda_k). \tag{A13}$$

As the full conditional distribution of E_k is bounded above, I can generate random draws from it using rejection sampling. Starting with a posterior sample $[\tau_k^{(p)}, \Upsilon^{(p)}, \Lambda_k^{(p)}]$, I draw one replicate $E_k^{(p)}$ for each $p = 1, \ldots, P$. For each p, I first generate E^* from a $\mathrm{Poisson}(\Lambda_k^{(p)})$ distribution and U from the uniform distribution. Then, if $U \leq (\mathbf{A}^{E^*})_{uv}/(\mathbf{P})_{uv}$, where $\Upsilon^{(p)} = u$ and $\tau_k^{(p)} = v$, I set $E_k^{(p)} = E^*$. Otherwise, I reject the current proposal and begin again by regenerating (E^*, U).

MCMC performance: I run my MCMC chains for 1.1×10^5 outer Metropolis-within-Gibbs cycles, discard the first 10^4 cycles as burn-in, and subsample every 10 cycles. This process retains $P = 10^4$ posterior samples with decreased autocorrelation. The total chain length and burn-in time appear moderately longer than required by examining time-series plots of the model log-likelihood during simulation.

To assess the performance of the MCMC sampler, I employ scaled SRQ plots (Mykland *et al.* 1995; Li *et al.* 2000; Suchard *et al.* 2002). SRQ plots are useful to demonstrate adequate sampler mixing within discrete model parameters. For the primary measures in this study, two important discrete parameters are the species tree Υ and the model mixture parameter ψ. In particular, I use SRQ plots to assess mixing when comparing the relative probabilities of two possible species trees and of differing stochastic models for HGT. In these SRQ plots, the local slope around a given point depicts the ratio of the relative posterior probability estimate based on the entire MCMC chain to an estimate based on a short segment of the chain around that point. Substantial deviation of the slope from one implies that the sampler is slowly mixing and, as a result, the chain is not sufficiently long to generate stable estimates. For continuous model parameters and Bayes factors based on the Savage-Dickey ratio, I assess convergence by comparing posterior estimates obtained from simulations of at least five independent chains with starting values drawn directly from the model priors.