Two Stationary Nonhomogeneous Markov Models of Nucleotide Sequence Evolution

VIVEK JAYASWAL1,2, LARS S. JERMIIN2,3,4,5,6,*, LEON POLADIAN1,2, AND JOHN ROBINSON1,2

1School of Mathematics and Statistics, 2Centre for Mathematical Biology, 3School of Biological Sciences, and 4Sydney Bioinformatics, University of Sydney, Sydney, NSW 2006, Australia; 5Division of Ecosystem Sciences at the Commonwealth Scientific Industrial and Research Organization, Canberra, ACT 2601, Australia; and 6Department of Biochemical Sciences “Rossi Fanelli”, University of Rome, “Sapienza” 00185 Rome, Italy.

*Correspondence to be sent to: Division of Ecosystem Sciences at the Commonwealth Scientific Industrial and Research Organization, Canberra, ACT 2601, Australia; E-mail: lars.jermiin@csiro.au.

Received 15 December 2009; reviews returned 16 March 2010; accepted 15 September 2010
Associate Editor: Marc Suchard

Abstract.—The general Markov model (GMM) of nucleotide substitution does not assume the evolutionary process to be stationary, reversible, or homogeneous. The GMM can be simplified by assuming the evolutionary process to be stationary. A stationary GMM is appropriate for analyses of phylogenetic data sets that are compositionally homogeneous; a data set is considered to be compositionally homogeneous if a statistical test does not detect significant differences in the marginal distributions of the sequences. Though the general time-reversible (GTR) model assumes stationarity, it also assumes reversibility and homogeneity. We propose two new stationary and nonhomogeneous models—one constrains the GMM to be reversible, whereas the other does not. The two models, coupled with the GTR model, comprise a set of nested models that can be used to test the assumptions of reversibility and homogeneity for stationary processes. The two models are extended to incorporate invariant sites and used to analyze a seven-taxon hominoid data set that displays compositional homogeneity. We show that within the class of stationary models, a nonhomogeneous model fits the hominoid data better than the GTR model. We note that if one considers a wider set of models that are not constrained to be stationary, then an even better fit can be obtained for the hominoid data. However, the methods for reducing model complexity from an extremely large set of nonstationary models are yet to be developed. [Invariable sites; maximum likelihood; nucleotide substitution model; phylogenetics; reversible nonhomogeneous model; stationary nonhomogeneous model.]

Most likelihood-based phylogenetic methods assume that evolutionary relationships can be represented by tree-like patterns of descent. The methods also assume that evolutionary processes leading to the accumulation of substitutions in diverging nucleotide sequences can be approximated by independent Markov processes, that is, processes in which the conditional rates of change at homologous sites depend only on the current states and are independent of the previous states. The Markov models that are used in these methods have many convenient properties and, depending on their design, may apply to stationary or nonstationary, reversible or nonreversible, and homogeneous or nonhomogeneous processes. In the context of a nucleotide sequence evolving along an edge of a phylogenetic tree, the evolutionary process is: stationary when the marginal distribution of the process does not change over that edge; reversible when the probability of sampling nucleotide i from the stationary distribution and going to nucleotide j is equal to that of sampling nucleotide j from the stationary distribution and going to nucleotide i; and homogeneous when the conditional rates of change are constant in time (for further details, see Jayaswal et al. 2005; Ababneh et al. 2006).

An evolutionary process has properties that are either local (pertain to a specific edge of the evolutionary tree) or global (pertain to the entire tree). For example, if the Markov process over an edge of the tree has fixed rates of substitution and is the same for all the edges, then the process is not only locally homogeneous but also globally homogeneous. Evolutionary processes may be partitioned into 8 scenarios depending on whether the data meet one or more of the stationary, reversible, and homogeneous conditions (Jermiin et al. 2008). However, because a reversible process by definition (Kolmogoroff 1936) is also a stationary process, only 6 of these scenarios are possible (Table 1).

Most model-based phylogenetic methods, that is, those implementing the general time-reversible (GTR) model by Lanave et al. (1984) and the special cases of this model, assume that the sequences have evolved under globally stationary, reversible, and homogeneous conditions (Scenario 1). A number of alternative phylogenetic methods have been developed to accommodate that diverging lineages may have evolved under different conditions. Although several of these methods relax the assumptions of global stationarity and homogeneity (e.g., Yang and Roberts 1995; Galtier and Gouy 1998; Galtier et al. 1999; Foster 2004; Dutheil and Boussau 2008), they still assume local homogeneity. By contrast, Barry and Hartigan’s (1987) model, henceforth called the BH model, is at the other end of the spectrum and suitable for nucleotide sequences that have evolved under locally nonstationary, nonreversible, and nonhomogeneous conditions (Jayaswal et al. 2005).

Jayaswal et al. (2007) extended the BH model to cases where sites may be divided into potentially variable and invariant sites. An invariant site is one where nucleotide substitutions cannot occur due to selective constraints; all other sites are potentially variable sites. The extended BH model provides a good fit between the data and the model when the nucleotide sequences have evolved under locally nonstationary, nonreversible,
nonhomogeneous conditions and the sites can be divided into potentially variable and invariant sites.

The extended BH model is the most general Markov model (GMM) under the assumption that processes on potentially variable sites are independent and identically distributed. However, this model may be too complex and have more parameters than are required for a particular data set. This is especially true for phylogenetic data sets that appear compositionally homogeneous, that is, data sets where Stuart’s (1955) matched-pairs test of marginal symmetry does not reject the null hypothesis of homogeneous marginal distributions. Because such data sets exist (e.g., the Xq13.3 data set from Weiss and von Haeseler 2003), it is desirable to simplify the extended BH model such that the Markov process is globally stationary but not necessarily reversible or homogeneous.

Here, we present two such Markov models—one model assumes the evolutionary process to be globally stationary, nonreversible and nonhomogeneous (Scenario 7) and the other model assumes the evolutionary process to be globally stationary, reversible and nonhomogeneous (Scenario 5). The two models, henceforth called the stationary Barry and Hartigan (SBH) and reversible Barry and Hartigan (RBH) models, are then extended to accommodate sites that are either potentially variable or invariant. This is followed by an analysis of data simulated on a five-taxon tree to show that the parameter estimates are accurate when the models are correct. Finally, we illustrate the merits of the new models by analyzing a seven-taxon hominoid data set.

**METHODS**

We represent the nucleotide sequence data set as a $K \times N$ matrix, where $K$ denotes the number of sequences (taxa) and $N$ denotes the number of nucleotides. Each column in this matrix represents a site in the sequence alignment, so the data set has $N$ sites. This matrix is used to obtain an unrooted binary tree $T$ with $K$ leaf nodes, $K - 2$ internal nodes, and $2K - 3$ edges.

We denote the set of leaf nodes by $L = \{1, \ldots, K\}$ and the set of internal nodes by $I = \{1, \ldots, I_{K-2}\}$. Let $V = L \cup I$ denote the set of all nodes and $E = \{(a, b) : a, b \in V \text{ and adjacent}\}$ denote the set of all edges. If an edge $(a, b)$ of the unrooted tree is deleted, then two rooted subtrees $T_{(a,b)}$ and $T_{(b,a)}$ are formed with roots at $a$ and $b$, respectively. Let $L_{(a,b)}$ denote the leaf nodes in subtree $T_{(a,b)}$ and let $L_{(b,a)}$ denote the leaf nodes in subtree $T_{(b,a)}$.

Let $B = \{A, C, G, T\}$ and $X_a$ and $X_b$ denote the random variables corresponding to the nucleotides at nodes $a$ and $b$ of the edge $(a, b)$. Let $x_a$ and $x_b$ denote the actual nucleotides at the nodes $a$ and $b$ of the edge $(a, b).$ Let $Q_{(a,b)}$ denote the $4 \times 4$ joint probability matrix such that

$$Q_{(a,b)}(x_a, x_b) = P(X_a = x_a, X_b = x_b); \ x_a \in B, \ x_b \in B. \quad (1)$$

Let $Q_a$ denote the vector of marginal probabilities at node $a$ such that

$$Q_a(x_a) = \sum_{x_b \in B} Q_{(a,b)}(x_a, x_b); \ (a, b) \in E. \quad (2)$$

We assume that the process acting on each site is Markovian and that the processes at different sites are independent and identically distributed. Also, for a specific pair of nucleotides at nodes $a$ and $b$, $(X_a, X_b) = (x_a, x_b)$, the conditional distributions of the nucleotides at the leaf nodes $L_{(a,b)}$ and $L_{(b,a)}$ are independent.

For a given site, $i$, if node $a$ is a leaf node, then $B_{ai} = \{x_a\}$, otherwise $B_{ai} = \{A, C, G, T\}$. The likelihood of nucleotides at leaf nodes for the $i$th site is given by

$$L_i = \sum_{x_a \in B_{ai}} \sum_{x_b \in B_{bi}} Q_{(a,b)}(x_a, x_b)P(L_{(a,b)}|x_a)P(L_{(b,a)}|x_b), \quad (3)$$

where the summations are only over the values in the sets $B_{ai}$ and $B_{bi}$, and

$$P(L_{(a,b)}|x_a) = \frac{Q_{L_{(a,b)}}(x_a)}{Q_a(x_a)}. \quad (4)$$

Here, $x_{L_{(a,b)}}$ denotes the vector of nucleotides at the $i$th site for the leaf nodes in $L_{(a,b)}$. Therefore, $Q_{L_{(a,b)}}(x_a, x_b)$ denotes the joint probability of $x_{L_{(a,b)}}$ and the nucleotide at node $a$.

Because the marginal probability at an internal node is consistent (Jayaswal et al. 2005), the ratio in Equation (4) does not depend on the term $Q_{(a,b)}(x_a, x_b)$ in Equation (3). Differentiating Equation (3) with respect to $Q_{(a,b)}(x_a, x_b)$, we obtain

$$L_i'(x_a, x_b) = \frac{\partial L_i}{\partial Q_{(a,b)}(x_a, x_b)} = \begin{cases} P(L_{(a,b)}|x_a)P(L_{(b,a)}|x_b), & \text{if } x_a \in B_{ai}, x_b \in B_{bi}, \\ 0, & \text{otherwise} \end{cases}$$

Now the log-likelihood for all the $N$ sites is given by

$$\log L = \sum_{i=1}^{N} \log L_i.$$

**TABLE 1.** Eight scenarios based on whether the evolutionary process is globally stationary and/or globally reversible and/or globally homogeneous

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Stationary</th>
<th>Reversible</th>
<th>Homogeneous</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Possible</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>By definition, impossible</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Possible</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Possible</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Possible</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>By definition, impossible</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Possible</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Possible</td>
</tr>
</tbody>
</table>
We use the expectation–maximization (EM) approach (Dempster et al. 1977) to estimate the proportion and nucleotide composition of invariable sites and to show that the Q-matrices can be estimated by maximizing the log-likelihood over the potentially variable sites. The constraints of stationarity or reversibility on the Q-matrices are incorporated using Lagrange multipliers (described below) and the maximization of the log-likelihood yields a set of simultaneous equations, which are solved using an iterative process similar to that first described by Barry and Hartigan (1987). The equations used in our iterative algorithm were designed such that if all the initial parameter estimates are positive and consistent with the relevant constraints, then at each step of the iteration, the parameter estimates, including the estimates of Lagrange multipliers, will remain positive and consistent with the constraints. Furthermore, empirical observations have revealed that the iterations converge to a common local maximum (Jayaswal et al. 2005), provided that the diagonal entries of the initial Q-matrices are larger than the off-diagonal entries. Choosing initial Q-matrices such that the off-diagonal entries are larger than the diagonal entries may result in convergence to a maximum lower than the common local maximum.

SBH and RBH Models

The two models assume that all the N sites are potentially variable, that sites evolve under independent and identically distributed Markov processes, and that the Markov process acting on a site is stationary.

Estimating Q-matrices and Lagrange multipliers for the SBH model.—Let \( \pi = \{ \pi_A, \pi_C, \pi_G, \pi_T \} \) be the vector of stationary probabilities. For the process to be stationary, the Q-matrix along an edge \((a, b)\) should satisfy the following constraints of marginal symmetry:

\[
\sum_{x_b \in B} Q_{(a,b)}(x_a, x_b) = \pi_{x_a} \tag{6}
\]

and

\[
\sum_{x_a \in B} Q_{(a,b)}(x_a, x_b) = \pi_{x_b}. \tag{7}
\]

Let \( \lambda = \{ \lambda_A, \lambda_C, \lambda_G, \lambda_T \} \) be the vector of Lagrange multipliers corresponding to the four constraints represented by Equation (6) and \( \delta = \{ \delta_A, \delta_C, \delta_G, \delta_T \} \) be the vector of Lagrange multipliers corresponding to the four constraints represented by Equation (7). Now the function to be maximized is

\[
\log L + \sum_{x_b \in B} \lambda_{x_b} \left( \pi_{x_b} - \sum_{x_a \in B} Q_{(a,b)}(x_a, x_b) \right) + \sum_{x_a \in B} \delta_{x_a} \left( \pi_{x_a} - \sum_{x_b \in B} Q_{(a,b)}(x_a, x_b) \right). \tag{8}
\]

Differentiating Equation (8) with respect to \( Q_{(a,b)}(x_a, x_b) \) and equating the partial derivatives to 0, we obtain

\[
\sum_{i=1}^{N} \frac{\partial \log L_i}{\partial Q_{(a,b)}(x_a, x_b)} = \lambda_{x_a} + \delta_{x_b}. \tag{9}
\]

We can use this condition to construct an iterative equation by multiplying both sides by \( Q_{(a,b)}(x_a, x_b) \) and rearranging it to give

\[
Q_{(a,b)}^\text{new}(x_a, x_b) = \frac{Q_{(a,b)}(x_a, x_b)}{(\lambda_{x_a} + \delta_{x_b})} \left[ \sum_{i=1}^{N} \frac{L_i(x_a, x_b)}{L_i} \right]. \tag{10}
\]

The latest estimates of all the parameters are used in the expression on the right and the expression on the left is interpreted as the new value of \( Q_{(a,b)}(x_a, x_b) \). The Lagrange multipliers used in Equation (10) are obtained using the relevant constraints as follows:

\[
\pi_{x_a} = \sum_{x_b \in B} Q_{(a,b)}(x_a, x_b) = \sum_{x_b \in B} \frac{Q_{(a,b)}(x_a, x_b)}{\lambda_{x_a} + \delta_{x_b}} \left[ \sum_{i=1}^{N} \frac{L_i(x_a, x_b)}{L_i} \right] \tag{11}
\]

or

\[
\lambda_{x_a}^\text{new} = \frac{\lambda_{x_a}}{\pi_{x_a}} \sum_{x_b \in B} Q_{(a,b)}(x_a, x_b) \left[ \sum_{i=1}^{N} \frac{L_i(x_a, x_b)}{L_i} \right], \tag{11}
\]

and

\[
\pi_{x_b} = \sum_{x_a \in B} Q_{(a,b)}(x_a, x_b) = \sum_{x_a \in B} \frac{Q_{(a,b)}(x_a, x_b)}{\lambda_{x_a} + \delta_{x_b}} \left[ \sum_{i=1}^{N} \frac{L_i(x_a, x_b)}{L_i} \right], \tag{12}
\]

or

\[
\delta_{x_b}^\text{new} = \frac{\delta_{x_b}}{\pi_{x_b}} \sum_{x_a \in B} Q_{(a,b)}(x_a, x_b) \left[ \sum_{i=1}^{N} \frac{L_i(x_a, x_b)}{L_i} \right]. \tag{12}
\]

Estimating Q-matrices and Lagrange multipliers for the RBH model.—The RBH model is obtained from the SBH model by adding the constraint of reversibility along each edge. A Markov process operating along an edge \((a, b)\) is reversible if the probability of starting at nucleotide \(e\) at node \(a\) and going to nucleotide \(f\) at node \(b\) is the same as the probability of starting at nucleotide \(f\) at node \(b\) and going to nucleotide \(e\) at node \(a\), that is, \( \pi_{e} P(f|e) = \pi_{f} P(e|f) \). This implies that \( Q_{(a,b)}(X_a = e, X_b = f) = Q_{(a,b)}(X_a = f, X_b = e) \) and, hence, that the Q-matrix for a reversible Markov process is fully symmetric. Therefore, instead of 8 Lagrange multipliers, we need only 4 Lagrange multipliers for the constrained optimization and

\[
Q_{(a,b)}^\text{new}(x_a, x_b) = \frac{Q_{(a,b)}(x_a, x_b)}{(\lambda_{x_a} + \lambda_{x_b})} \sum_{i=1}^{N} \left[ \frac{L_i(x_a, x_b)}{L_i} + \frac{L_i(x_b, x_a)}{L_i} \right]. \tag{13}
\]
It should be noted that Equation (13) preserves the symmetry of Q-matrices at each step provided the initial matrices were also symmetric. The Lagrange multipliers used in Equation (13) are obtained by using the relevant constraints as follows:

$$\pi_{x_i} = \sum_{x_j \in B} Q_{(a,b)}^{\text{new}}(x_i, x_j),$$

or

$$\lambda_{x_i}^{\text{new}} = \frac{\lambda_{x_i}}{\pi_{x_i}^{\text{new}}} \sum_{x_j \in B} Q_{(a,b)}^{\text{new}}(x_i, x_j).$$

(14)

Because the SBH and RBH models assume global stationarity, the marginal probabilities of all the Q-matrices are set to $\pi_i$. Therefore, the SBH and RBH models always satisfy the constraint of internal consistency. Estimating the stationary probabilities.—For a particular site, $i$, the likelihood given by Equation (3) can be rewritten as follows:

$$L_i = \sum_{x_i \in B} \pi_{x_i} \sum_{x_j \in B} P_{(a,b)}(x_i | x_j)P_{(b,b)}(x_j | x_a)P_{(a,a)}(x_a | x_b),$$

(15)

where $\pi_{x_i}$ denotes the stationary probability of nucleotide $x_i$, and $P_{(a,b)}(x_i | x_j)$ denotes the conditional probability of the nucleotide at node $b$ being $x_j$, given that the nucleotide at node $a$ is $x_i$. The new estimate of $\pi_i$ can be obtained by maximizing the function

$$\log L + \epsilon \left(1 - \sum_{j \in B} \pi_j \right),$$

(16)

where $\epsilon$ is a Lagrange multiplier. Differentiating Equation (16) with respect to $\pi_i$ and equating the partial derivatives to 0, we obtain

$$\pi_{x_i}^{\text{new}} = \frac{\pi_{x_i}}{\epsilon} \sum_{i=1}^{N} I(x_i = j) L_i'(x_i, x_j)$$

(17)

where $I(x_i = j)$ is an indicator function that takes the value 1 when the nucleotide at node $a$ for the $i$th site is $j$ and 0, otherwise. The value for $\epsilon$ can be obtained as follows:

$$\epsilon = \sum_{j \in B} \pi_{x_i}^{\text{new}} \left(1 - \sum_{i=1}^{N} I(x_i = j) L_i'(x_i, x_j) \right).$$

(18)

The algorithms for estimating model parameters of SBH and RBH are described in Appendix 1. It should be noted that the maximum-likelihood estimates of Q-matrices are obtained such that the marginal probability vectors for the Q-matrices are equal to $\pi_i$, a prespecified vector of stationary probabilities. A change in stationary probabilities from $\pi$ to $\pi^{\text{new}}$ does not result in an implicit change in the marginal probabilities of the Q-matrices. Therefore, the Q-matrices and $\pi$ cannot be estimated simultaneously.

Relationship Between Models Based on Joint Probability Matrices and Models Based on Rate Matrices

In our models, $Q_{(a,b)}$ represents the matrix of joint probabilities at nodes $a$ and $b$. For the SBH model,

$$Q_{(a,b)} = Q_{(b,a)}^{\text{T}} = \pi,$$

(19)

where the vector of stationary probabilities, $\pi$, is the same for all Q-matrices, and 1 denotes the vector $[1 1 1]^T$. For the RBH model, $Q_{(a,b)} = Q_{(b,a)}$, so Condition (19) is implicitly satisfied.

Let $\Pi = \text{diag}(\pi)$ denote the diagonal matrix of stationary probabilities and let $P_{(a,b)} = \Pi^{-1} Q_{(a,b)}$ denote the conditional probability matrix over an edge $(a, b)$. If we assume that there is a locally homogeneous Markov process with rate matrix $R$ over the edge $(a, b)$, then

$$P_{(a,b)} = P(t) = e^{Rt},$$

where $t$ denotes the length of the edge $(a, b)$. The rate matrix must have the property that $R1 = 0$ and that the off-diagonal elements are nonnegative. For example, a time-reversible rate matrix is of the form

$$R = \begin{bmatrix}
-\rho_a & \rho_b & \rho_c & \rho_d & \rho_e & \rho_f \\
\rho_a & -\rho_b & \rho_c & \rho_d & \rho_e & \rho_f \\
\rho_b & \rho_d & -\rho_f & \rho_e & \rho_d & \rho_e \\
\rho_c & \rho_e & \rho_f & -\rho_e & \rho_f & \rho_e \\
\rho_d & \rho_e & \rho_f & \rho_e & -\rho_f & \rho_e \\
\rho_e & \rho_f & \rho_e & \rho_f & \rho_e & -\rho_f
\end{bmatrix} = \Pi R,$$

(20)

where the elements $\rho_a, \ldots, \rho_f$ are nonnegative.

In general, for a joint probability matrix $Q_{(a,b)}$, the corresponding conditional probability matrix $P_{(a,b)}$ is not always associated with a rate matrix of the form described above. This is the embedding problem for Markov chains considered by Kingman (1962).

Let the term “rate matrix–based model” denote a model with a homogeneous Markov process along each edge. Then, $P_{(a,b)}$ is obtained explicitly using $e^{Rt}$ (not $\Pi^{-1} Q_{(a,b)}$). If there is a distinct R-matrix for each edge such that every R-matrix satisfies Equation (20) and $\Pi$ is the same for all edges, then the rate matrix–based model will have exactly the same number of free parameters as the RBH model. However, such a rate matrix–based model assumes local homogeneity ($R$ is fixed over an edge) and is a restricted case of the RBH model. Similarly, a rate matrix–based model that has the same $\Pi$ for all the edges and assumes local homogeneity but not local reversibility is a restricted case of the SBH model. Therefore, the rate matrix–based models will return log-likelihoods that are lower than those returned using our models.

The necessary and sufficient conditions for a $4 \times 4$ conditional probability matrix ($P_{(a,b)}$) to have an associated rate matrix ($R$) are unknown. Therefore, a comparison of the log-likelihoods obtained under the two models is necessary to test the assumption of local homogeneity.

SBH + I2 and RBH + I2 Models

So far, we have assumed that every site in the sequence alignment is potentially variable. We will now
consider extensions to the SBH and RBH models, namely the SBH + 1 and RBH + 1 models, which relax this assumption. The new models assume that a site is either potentially variable or invariable that all potentially variable sites evolve under independent and identically distributed Markov processes and that the Markov process acting on a potentially variable site is stationary. In this paper, 1 represents a model that does not assume the marginal distributions of the potentially variable sites and the invariable sites to be equal, whereas 1 represents a model that assumes the two marginal distributions to be equal.

Estimation of parameters for invariable sites.—If a site has the same nucleotide in all the K sequences, then it is referred to as a constant site. Therefore, constant sites comprise all the invariable sites and those potentially variable sites that are observed to have the same nucleotide in all the K sequences. Let \( \alpha \) and \( \beta \) denote the probability of a site being potentially variable or invariable, respectively; it then follows that \( \alpha + \beta = 1 \). Let \( \mathcal{H} \) denote the set of column vectors of length \( K \), with all elements of a vector being identical and equal to A, C, G, or T. Thus, \( \mathcal{H} \) denotes the set of patterns corresponding to constant sites. Let \( \pi_{\text{inv}} \) denote the probability that the site is of type \( h \in \mathcal{H} \), given that it is an invariable site and \( \pi_{\text{inv}} = \{ \pi_{A}^{\text{inv}}, \pi_{C}^{\text{inv}}, \pi_{G}^{\text{inv}}, \pi_{T}^{\text{inv}} \} \). Let \( L_h \) denote the probability that the site is of type \( h \in \mathcal{H} \), given that it is a potentially variable site. Let \( P(h) \) denote the probability of a constant site of type \( h \in \mathcal{H} \).

Using the EM approach described in Appendix 2,

\[
P(h) = \alpha L_h + \beta \pi_{h}^{\text{inv}}. \tag{21}
\]

Since \( \alpha = 1 - \beta \), Equation (21) can be summed over all possible values of \( h \) to obtain

\[
\beta = \frac{\sum_{h \in \mathcal{H}} [P(h) - L_h]}{1 - \sum_{h \in \mathcal{H}} L_h}. \tag{22}
\]

Having obtained an estimate of \( \beta \) (and thus also of \( \alpha \)), Equation (21) can be used to estimate \( \pi_{h}^{\text{inv}} \) as follows:

\[
\pi_{h}^{\text{inv}} = \frac{P(h) - \alpha L_h}{\beta}. \tag{23}
\]

Total log-likelihood.—For the SBH and RBH models, the log-likelihood was obtained using Equation (5) because every site was potentially variable. For the SBH + 1 and RBH + 1 models, this log-likelihood has to be modified to incorporate both potentially variable and invariable sites. Let \( N_1 \) denote the number of constant sites and let \( N = N_1 - N_2 \) denote the remaining sites. Because the sites are assumed to be independent, we can rearrange them without loss of generality such that the last \( N_2 \) sites in the alignment correspond to constant sites. Now the total log-likelihood is

\[
\log L = \sum_{i=1}^{N_1} \log(\alpha L_i) + \sum_{i=N_1+1}^{N} I(\text{pat}_i = h) \log(\alpha L_h + \beta \pi_{h}^{\text{inv}}), \tag{24}
\]

where \( L_i \) denotes the likelihood of the pattern at site \( i \), given that it is potentially variable, and \( I(\text{pat}_i = h) \) is an indicator function that takes the value 1 if the pattern at site \( i \) is equal to \( h \) and 0, otherwise.

The invariable sites could be estimated using Lagrange multipliers for \( \alpha \), \( \beta \), and \( \pi_{h}^{\text{inv}} \) (e.g., Jayaswal et al. 2007). However, the EM approach described in this section has the advantage that it does not require any modification to the existing formulae for estimating the Q-matrices. This approach can in fact be used to estimate the invariable sites’ parameters for any maximum-likelihood–based model without the need for model-specific equations.

Free Parameters for the New Models

SBH model.—Each Q-matrix has 9 free parameters since Equations (6) and (7) have to be satisfied. In addition, there are 3 free parameters for \( \pi \), the vector of stationary probabilities. Therefore, the SBH model has \( 18K - 24 \) free parameters, where \( K \) denotes the number of sequences.

RBH model.—Each Q-matrix has 6 free parameters since the matrix is fully symmetric. In addition, there are 3 free parameters for \( \pi \). Therefore, the RBH model has \( 12K - 15 \) free parameters.

For the SBH + 1 and RBH + 1 models, there are 4 additional free parameters since \( \alpha + \beta = 1 \) and \( \sum_{X \in \{A,C,G,T\}} \pi_{X}^{\text{inv}} = 1 \).

RESULTS AND DISCUSSION

In this section, we compare the log-likelihoods obtained under different models when the sequences have evolved under stationary and nonreversible conditions or stationary and reversible conditions. The process is not assumed to be locally homogeneous. Next, we determine the accuracy of estimates obtained under the SBH + 1 and RBH + 1 models. Finally, we compare the results obtained using different models for a seven-taxon hominoid data set. It should be noted that all the log-likelihood values mentioned in this section are at the maximum-likelihood estimates.

Model Comparison

We compare the log-likelihoods obtained under the GTR, RBH, SBH, and BH models when the SBH model is correct or the RBH model is correct. There are two possible approaches for obtaining Q-matrices for simulation that satisfy the model-dependent constraints.
of stationarity and reversibility. One approach requires the use of a random number generator for obtaining the Q-matrices with the relevant constraints. The second approach requires the analysis of an arbitrary data set using the appropriate model, thereby, obtaining Q-matrices with the relevant constraints. We used the latter approach and performed the following steps to generate the Q-matrices:

1. Analyzed a bacterial data set (Jayaswal et al. 2007) using the BH model. A real biological data set was chosen to obtain parameters under realistic conditions.
2. Used the estimates of parameters obtained in Step 1 to simulate a five-taxa data set of length 4000 sites.
3. Estimated the Q-matrices for the simulated data set (generated in Step 2) under the RBH and SBH models, thereby obtaining $Q_{RBH}$ and $Q_{SBH}$, respectively, along each edge. The presence of a large number of sites (4000) and hence a large number of distinct patterns in the simulated data set ensured that none of the elements in a $Q_{RBH}$ or $Q_{SBH}$ matrix was close to 0 and, along an edge $(a, b)$, the joint probability of each of the 16 possible combinations of nucleotides (AA, ..., TT) was strictly positive.

The $Q_{RBH}$ or $Q_{SBH}$ matrices along the edges were sufficiently different to ensure that the evolutionary process over the tree was distinguishable from a globally homogeneous process.

Case 1: SBH is the true model.—Using the $Q_{SBH}$ matrices, we generated sequences on a five-taxon tree. We considered sequences of 1200, 5000, and 10,000 nucleotides and simulated 1000 data sets for each sequence length. For a five-taxa tree, the number of free parameters under the GTR, RBH, SBH, and BH models are 15, 45, 66, and 87, respectively. Because the four models are nested, we used a likelihood-ratio test (LRT) (Neyman and Pearson 1933) to identify the best model, that is, the simplest model consistent with the data. The models were compared sequentially in increasing order of complexity, starting with the simple GTR model and moving toward the parameter-rich BH model. Table 2a shows that, as expected, the improvement in log-likelihood is significant for the comparisons of GTR versus RBH and RBH versus SBH models with $P(\chi^2_{AD} \geq 2\Delta\log L) \approx 0$, where $\Delta D$ denotes the difference in the number of free parameters and $\Delta\log L$ denotes the difference in log-likelihoods. At the same time, the improvement in log-likelihood as one moves from the SBH model to the BH model is not significant and is due to overparameterization.

We observed that $\Delta\log L = \log L_{RBH} - \log L_{GTR}$ and $\Delta\log L = \log L_{SBH} - \log L_{RBH}$ increase with an increase in the sequence length (Table 2a). However, in molecular phylogenetics, sequence of 5000 and 10,000 nucleotides are unusual for a single locus (for an exception, see the Xq13.3 data from Weiss and von Haeseler 2003); typical loci used in phylogenetics contain approximately 1200 nucleotides (e.g., the mitochondrially encoded Cytochrome b gene). Because the improvement is significant even for a short sequence length, the SBH model should be preferred over the GTR and RBH models when the data have evolved under stationary but not locally homogeneous or reversible conditions.

The Akaike Information Criterion (AIC; Akaike 1974; Burnham and Anderson 2004) also favored the SBH model over the other models for all the simulated data sets (results not shown). Therefore, the results from the LRT and the AIC were in agreement.

Case 2: RBH is the true model.—Using the $Q_{RBH}$ matrices, we generated sequences on a five-taxon tree. We considered three different sequence lengths (same as in Case 1) and the results in Table 2b show that the improvement in log-likelihood as one moves from the GTR to RBH model is significant, whereas the

| Table 2. Difference in log-likelihoods obtained under different models |
|---------------------|-------------------|-------------------|-------------------|-------------------|
| **a. SBH is the correct model** | **RBH versus GTR** | **SBH versus RBH** | **BH versus SBH** |
| Sequence length | $\Delta D$ | $\Delta\log L^a$ | $p^b$ | $\Delta D$ | $\Delta\log L^a$ | $p^b$ | $\Delta D$ | $\Delta\log L^a$ | $p^b$ |
| 1200 | 30 | 77.20 ± 11.80 | $\approx 0$ | 21 | 63.33 ± 9.56 | $\approx 0$ | 21 | 10.63 ± 3.35 | 0.44 |
| 5000 | 30 | 283.12 ± 22.96 | $\approx 0$ | 21 | 231.14 ± 18.79 | $\approx 0$ | 21 | 10.66 ± 3.27 | 0.44 |
| 10,000 | 30 | 550.21 ± 32.25 | $\approx 0$ | 21 | 453.64 ± 29.14 | $\approx 0$ | 21 | 10.53 ± 3.35 | 0.46 |
| **b. RBH is the correct model** | **RBH versus GTR** | **SBH versus RBH** |
| Sequence length | $\Delta D$ | $\Delta\log L^a$ | $p^b$ | $\Delta D$ | $\Delta\log L^a$ | $p^b$ |
| 1200 | 30 | 46.67 ± 8.71 | $\approx 0$ | 21 | 9.43 ± 3.01 | 0.59 |
| 5000 | 30 | 158.33 ± 15.05 | $\approx 0$ | 21 | 9.95 ± 3.26 | 0.53 |
| 10,000 | 30 | 307.73 ± 20.63 | $\approx 0$ | 21 | 9.62 ± 3.01 | 0.57 |

$^a$Values after ± indicate sample standard deviation.

$^b p = P(\chi^2_{AD} \geq 2\Delta\log L)$. 

Downloaded from https://academic.oup.com/sysbio/article/60/1/74/1652292 by guest on 12 March 2022
improvement in log-likelihood as one moves from the RBH to SBH model is not significant and is due to overparameterization.

For simulated data sets of length 1200, the AIC favored the RBH model in 97.1% of the cases and the simpler (but incorrect) GTR model in 2.9% of the cases. For simulated data sets of length 5000, the AIC favored the RBH model in 99.5% of the cases and the more complex (but incorrect) SBH model in the remaining cases. For simulated data sets of length 10,000, the AIC favored the RBH model in all but one case. This shows that for short sequence lengths, the LRT and AIC can occasionally give conflicting results. In our simulation studies, the LRT always picked the true model as the best model, unlike the AIC that occasionally picked the wrong model as the best model.

Accuracy of Parameter Estimates

In this section, we compare the expected and observed estimates of the invariable site parameters and \( \pi \) for potentially variable sites under the SBH + I\(_2\) and RBH + I\(_2\) models.

Case 1: SBH + I\(_2\) model.—In order to determine the accuracy of parameters obtained under the SBH + I\(_2\) model, we simulated 1000 data sets with 1200 sites being potentially variable and 600 sites being invariable. Because the SBH + I\(_2\) model does not require the nucleotide composition of the potentially variable sites to be the same as that for invariable sites, we considered different values of \( \pi \) for potentially variable and invariable sites. The results in Table 3 show that the parameter estimates are accurate when the model is correct.

Case 2: RBH + I\(_2\) model.—As for the SBH + I\(_2\) model, we simulated 1000 data sets with 1200 sites being potentially variable and 600 sites being invariable. The results in Table 4 show that the parameter estimates are accurate when the RBH + I\(_2\) model is correct.

Hominoid Data Set

We considered a seven-taxon hominoid data set comprising 1809 nucleotides from the mitochondrially encoded NADH dehydrogenase subunit 5 genes of Homo sapiens (Hsap), Pan troglodytes (Ptro), Pan paniscus (Ppan), Gorilla gorilla (Ggor), Pongo pygmaeus (Ppyg), and Macaca fascicularis (Mfasc). The data set was obtained from GenBank accession number NC_001643.

### Table 3. Parameter estimates under the SBH + I\(_2\) model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expected</th>
<th>Observed(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>0.33</td>
<td>0.32 ± 0.02</td>
</tr>
<tr>
<td>( \pi_{\text{inv}} )</td>
<td>0.33</td>
<td>0.34 ± 0.03</td>
</tr>
<tr>
<td>( \pi_A^{\text{inv}} )</td>
<td>0.26</td>
<td>0.24 ± 0.02</td>
</tr>
<tr>
<td>( \pi_T^{\text{inv}} )</td>
<td>0.17</td>
<td>0.15 ± 0.04</td>
</tr>
<tr>
<td>( \pi_{\text{ac}} )</td>
<td>0.25</td>
<td>0.26 ± 0.03</td>
</tr>
<tr>
<td>( \pi_A )</td>
<td>0.18</td>
<td>0.18 ± 0.02</td>
</tr>
<tr>
<td>( \pi_C )</td>
<td>0.31</td>
<td>0.31 ± 0.02</td>
</tr>
<tr>
<td>( \pi_T )</td>
<td>0.19</td>
<td>0.19 ± 0.02</td>
</tr>
</tbody>
</table>

\(^a\) Values after ± indicate sample standard deviation.

### Table 4. Parameter estimates under the RBH + I\(_2\) model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expected</th>
<th>Observed(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>0.33</td>
<td>0.33 ± 0.018</td>
</tr>
<tr>
<td>( \pi_{\text{inv}} )</td>
<td>0.33</td>
<td>0.33 ± 0.014</td>
</tr>
<tr>
<td>( \pi_A^{\text{inv}} )</td>
<td>0.25</td>
<td>0.26 ± 0.016</td>
</tr>
<tr>
<td>( \pi_T^{\text{inv}} )</td>
<td>0.17</td>
<td>0.17 ± 0.017</td>
</tr>
<tr>
<td>( \pi_A )</td>
<td>0.25</td>
<td>0.24 ± 0.012</td>
</tr>
<tr>
<td>( \pi_C )</td>
<td>0.17</td>
<td>0.18 ± 0.002</td>
</tr>
<tr>
<td>( \pi_T )</td>
<td>0.31</td>
<td>0.31 ± 0.004</td>
</tr>
</tbody>
</table>

\(^a\) Values after ± indicate sample standard deviation.

For a seven-taxon tree, there are 945 possible unrooted tree topologies. The tree topology shown in Figure 1 had the highest log-likelihood under each of the four models—GTR + I\(_1\), RBH + I\(_2\), SBH + I\(_2\), and BH + I\(_2\); we, henceforth, refer to this particular tree topology as T\(_{\text{best}}\). It should be noted that T\(_{\text{best}}\) differs from the assumed correct phylogenetic tree (T\(_{\text{ac}}\)) of hominoids, which has human as the sister taxon of chimpanzee and bonobo (e.g., Raauw et al. 2005). One possible reason for this could be that the sequences are relatively short and, therefore, that stochastic errors are preventing identification of the correct tree. The difference in
log-likelihoods of $T_{\text{best}}$ and $T_{\text{ac}}$ was found to be 2.41, 4.02, 4.03, and 5.10 for the GTR + $I_1$, RBH + $I_2$, SBH + $I_2$, and BH + $I_2$ models, respectively. We used the approximately unbiased test (Shimodaira 2002), which is implemented in CONSEL (Shimodaira and Hasegawa 2001), to test the null hypothesis that the tree topology set ($T_{\text{best}}, T_{\text{ac}}$) contains the true tree (i.e., the actual tree of evolution). We tested the null hypothesis for only the BH + $I_2$ model because it had the largest difference ($\log L_{T_{\text{best}}} - \log L_{T_{\text{ac}}} = 5.10$) and obtained $p$ values of 0.742 and 0.258, respectively, for $T_{\text{best}}$ and $T_{\text{ac}}$. These results show that either of the two trees could be the true tree and that the two trees are statistically indistinguishable.

Model comparisons.—The results in Table 6 show that the increase in log-likelihood as one moves from the GTR + $I_1$ to GTR + $I_2$ model is statistically significant ($p < 10^{-5}$; Table 6b). Therefore, a model that does not assume $\pi = \pi^\text{inv}$ should be preferred. Table 7 shows the estimates of $\beta$, $\pi^\text{inv}$, and $\pi$ for the GTR + $I_1$ and GTR + $I_2$ models, and as expected, $\pi^\text{inv}$ and $\pi$ are very different.

Because the LRT is applicable for nested models, we compared the GTR + $I_2$ model with the RBH + $I_2$ model and found some evidence that the GTR + $I_2$ model is not sufficient to explain the hominoid data set ($p = 0.019$; Table 6b). A comparison of RBH + $I_2$ and SBH + $I_2$ models showed that the RBH + $I_2$ model is sufficient to explain the data set ($p = 0.12$; Table 6b). Therefore, a comparison of the GTR + $I_2$ and RBH + $I_2$ stationary models indicated that the process was not globally homogeneous.

Within the class of stationary models, there is a range of increasingly complex models between the GTR + $I_2$ and the RBH + $I_2$ models. We denote these models as GTR$_G + I_2$ models, where $G$ represents the number of distinct rate matrices ($R_1, \ldots, R_G$) in the phylogenetic tree (see Equation 20 for the matrix structure). For the seven-taxon hominoid data set, if the tree is unrooted, then the maximum permissible value for $G$ is 11. The GTR$_{11} + I_2$ model has a distinct rate matrix per edge. In addition, if we assume the evolutionary process to be stationary, then all the edges have the same $\pi$. This stationary GTR$_{11} + I_2$ model has the same number of free parameters as the RBH + $I_2$ model. However, the stationary GTR$_{11} + I_2$ model assumes local homogeneity and is, therefore, a restricted case of the RBH + $I_2$ model. The stationary GTR$_{11} + I_2$ model returned a log-likelihood of $-3550.938$, which is slightly lower than $-3549.89$, the log-likelihood of the RBH + $I_2$ model (Table 6a). This is consistent with our remark in the previous section on rate matrices that the RBH model does not assume local homogeneity and, therefore, may return a higher log-likelihood compared with a model that assumes local (but not global) homogeneity. However, the small difference between the two models suggested that for the hominoid data set, the evolutionary process could be modeled by a locally homogeneous Markov process.

In general, for the GTR$_G + I_2$ models, the log-likelihood is dependent on the assignment of edges to the $G$ groups. For example, for a seven-taxon unrooted tree, there are 11 edges and if $G = 2$, the two rate matrices ($R_1$ and $R_2$) could be assigned to the edges in

![Figure 1. The most likely tree ($T_{\text{best}}$) of hominoids inferred using GTR + $I_1$, RBH + $I_2$, SBH + $I_2$, and BH + $I_2$ models](https://academic.oup.com/sysbio/article/60/1/74/1652292)

**Table 6.** Model comparison for the seven-taxon hominoid data set and tree topology $T_{\text{best}}$

<table>
<thead>
<tr>
<th>ID</th>
<th>Model</th>
<th>Log-likelihood</th>
<th>Free parameters</th>
<th>AIC</th>
<th>b. LRT results for the various nested models</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GTR + $I_1$</td>
<td>$-3599.72$</td>
<td>20</td>
<td>7239.44</td>
<td>2. versus 1</td>
</tr>
<tr>
<td>2</td>
<td>GTR + $I_2$</td>
<td>$-3586.41$</td>
<td>23</td>
<td>7218.81</td>
<td>4. versus 2</td>
</tr>
<tr>
<td>3</td>
<td>GTR + $I_2 + f_0$</td>
<td>$-3580.27$</td>
<td>27</td>
<td>7214.54</td>
<td>7. versus 2</td>
</tr>
<tr>
<td>4</td>
<td>GTR$_2$ + $I_2$ (Stationary)</td>
<td>$-3578.37$</td>
<td>28</td>
<td>7212.73</td>
<td>8. versus 7</td>
</tr>
<tr>
<td>5</td>
<td>GTR$_2$ + $I_2$ (Nonstationary)</td>
<td>$-3563.36$</td>
<td>31</td>
<td>7190.40</td>
<td>2. versus 1</td>
</tr>
<tr>
<td>6</td>
<td>GTR$_{11}$ + $I_2 + f_0$</td>
<td>$-3548.33$</td>
<td>60</td>
<td>7223.06</td>
<td>4. versus 2</td>
</tr>
<tr>
<td>7</td>
<td>RBH + $I_2$</td>
<td>$-3549.89$</td>
<td>73</td>
<td>7255.32</td>
<td>7. versus 2</td>
</tr>
<tr>
<td>8</td>
<td>SBH + $I_2$</td>
<td>$-3528.53$</td>
<td>106</td>
<td>7289.70</td>
<td>8. versus 7</td>
</tr>
<tr>
<td>9</td>
<td>BH + $I_2$</td>
<td>$-3509.53$</td>
<td>139</td>
<td>7333.57</td>
<td>2. versus 1</td>
</tr>
</tbody>
</table>

- Stationary GTR$_2$ + $I_2$ model.
- Nonstationary GTR$_2$ + $I_2$ model with $f_0 = \pi_2$.
- $S$-matrix is the same for all edges.
- Second-order bias corrected AIC value.
- $p = P(\chi^2_{AD} > 2\Delta \log L)$, where $\Delta D =$ difference in number of free parameters and $\Delta \log L =$ difference in log-likelihoods.
2^{10} - 1 different ways. Because it is infeasible to test all possible combinations, these models are difficult to use in practice unless a heuristic approach is implemented.

Because there is some evidence that the evolutionary process is not globally homogeneous, we considered the simplest extension of the GTR + I2 model, namely, the GTR2 + I2 model. We tried some of the 2^{10} - 1 combinations for assigning R1 or R2 to the 11 edges and obtained the log-likelihood of -3578.37 (Table 6a). We kept the rate matrices for the subtree (Human, Gorilla, [Bonobo, Chimpanzee]) the same owing to their relatively recent divergence. Let R1 denote the rate matrix for this subtree. For the remaining edges, we tried different combinations of R1 and R2 and found that the best log-likelihood was obtained when R1 was assigned to the subtree (Human, Gorilla, [Bonobo, Chimpanzee]) and all the remaining edges were assigned R2. Though we considered two different rate matrices, the same π was assigned to all the edges, thereby ensuring that the process was stationary.

Though the hominoid data set is compositionally homogeneous and its analysis using stationary models is appropriate, there may be more complex models that provide a better fit than the stationary models. Therefore, we obtained the log-likelihoods for a few cases where the evolutionary process was not stationary. This was achieved by considering a rooted tree such that the marginal probabilities at the root (f0) were different from those for the rate matrices along the edges. The root was placed on the edge linking Macaque and Gibbon. The simplest nonstationary model has the same rate matrix for all the edges but f0 is different from the π for the edges. The log-likelihood and AIC value for this GTR + I2 + f0 model were -3580.27 and 7214.54, respectively. The GTR + I2 model and the stationary GTR2 + I2 model were not nested and therefore could not be compared using the LRT. For the AIC, the stationary GTR2 + I2 model provided a slightly better fit than the nonstationary GTR + I2 + f0 model (7212.73 vs. 7214.54; Table 6a). One of the more complex nonstationary models has a different π per edge but the same S-matrix for all the edges. In addition, there is a distinct f0. This nonstationary model has 12 edges and is denoted as GTR12 + I2 + f0. It returned an AIC value of 7223.06 (Table 6a), which was higher than that observed for the stationary GTR2 + I2 model. Finally, we considered a nonstationary model that extended the stationary GTR2 + I2 model. Unlike the stationary GTR2 + I2 model that had R1 = S1Π and R2 = S2Π, the nonstationary model had R1 = S1Π1 and R2 = S2Π2. We set f0 = πf because this gave an AIC value lower than that for f0 = πf ≠ πf (result not shown). The nonstationary GTR2 + I2 model had the lowest AIC value (7190.40; Table 6a) among the set of models investigated by us. However, our search of the nonstationary and/or nonhomogeneous models was not exhaustive and another model may provide a better fit.

### Conclusion

We considered two Markov models of nucleotide sequence evolution that require the evolutionary process to be globally stationary but not homogeneous. Therefore, the SBH and RBH models are special cases of the BH model. For a K-taxon tree, the SBH model has 18K - 24 free parameters and the RBH model has 12K - 15 free parameters. The RBH and SBH models, therefore, have more parameters than the GTR model (2K + 5 free parameters) and less parameters than the BH model (24K - 33 free parameters). Consequently, in terms of complexity, the new models lie between the GTR model and the parameter-rich BH model and can be used to test the assumptions of stationarity, reversibility and homogeneity within the framework of nested models.

The RBH model considers the Markov process to be globally stationary and reversible but not homogeneous. In biological terms, a stationary but nonhomogeneous process represents an evolutionary process that has the same marginal probabilities for A, C, G, and T over the entire phylogenetic tree but the rate of evolution varies over the tree. Reversibility implies that the pairwise joint probabilities are identical along an edge (e.g., Q(a,b)(Xa = A, Xb = C) = Q(a,b)(Xa = C, Xb = A) over the edge (a,b)).

The SBH model is one where the biological process maintains the marginal probabilities of A, C, G, and T but does not require the pairwise joint probabilities to be the same (e.g., Q(a,b)(Xa = A, Xb = C) ≠ Q(a,b)(Xa = C, Xb = A)). This implies that the changes from A to C are not compensated by changes from C to A. Currently, we do not know which biological data sets satisfy this property, but our simulations show that the RBH model is insufficient to explain data sets that are generated under stationary but nonreversible conditions. We believe that the SBH model is a useful link between the RBH (stationary, reversible, and nonhomogeneous) model and BH (nonstationary, nonreversible, and nonhomogeneous) model.

The new models can be used to test the assumptions of stationarity and reversibility. If the difference in log-likelihoods between the RBH and SBH models is
If the difference in log-likelihoods between the SBH and BH models is statistically significant, then the process is not reversible. If the difference in log-likelihoods is not statistically significant, then the process is stationary or reversible. This test of stationarity differs from the matched-pairs test of marginal symmetry described by Ababneh et al. (2006). The latter test involves pairwise comparison of sequences and if any of the tests are statistically significant (after accounting for multiple comparisons), then there is evidence against the evolutionary process being stationary over the entire tree. However, Ababneh et al. (2006) did not consider the evolutionary processes over individual edges and, hence, cannot identify the actual edges over which the process is not stationary. In contrast, if the difference in log-likelihoods obtained under the SBH and BH models is significant, then the Q-matrices obtained under the BH model can be used to identify the edges over which the assumption of stationarity is not valid. However, one limitation of the SBH and RBH models could be the relatively large number of parameters to be estimated, and this might limit the number of taxa that can be analyzed using these parameter-rich models even when the GTR model is inappropriate.

The discovery that the GTR + $I_2$ model yields a significantly better fit than the GTR + $I_1$ model for the hominoid data set is interesting because it shows that it may not always be appropriate to assume that the marginal distribution of the variable sites is identical to that of the potentially variable sites. From the biological point of view, the result is easy to explain because we are dealing with two realistic classes of sites, one class being free to evolve and the other unable to evolve. Therefore, assuming that these two classes of sites should have the same marginal distribution is unrealistic. From a phylogenetic point of view, there are gains to be made by using the GTR + $I_2$ model. The advantages include a better fit of the model to the data and more accurate estimates of the relevant parameters. Therefore, it might be worth considering the GTR + $I_2$ model in future phylogenetic programs. A software implementation of our models can be downloaded from http://www.maths.usyd.edu.au/u/johnr or from under Publishing History at http://www.csiro.au/people/Lars.Jermiin.html.

**ACKNOWLEDGMENTS**

We wish to thank Bastien Boussau, Stephane Guindon, Marc Suchard, Jack Sullivan, and an anonymous reviewer for their many constructive comments.
these parameters is not crucial but we typically set \( \pi \) to the average nucleotide composition of all the taxon and the initial estimates of the Lagrange multipliers to half the number of sites.

1. For each edge of the phylogenetic tree,
   a. Obtain new estimates of Lagrange multipliers using the latest estimates of \( Q \)-matrices;
   b. Obtain new estimates of \( Q \)-matrices using the latest estimates of Lagrange multipliers;
   c. Calculate the residual
   \[
   \sum_{x_i \in B} \sum_{x_j \in B} [Q_{(a,b)}^{\text{new}}(x_i, x_j) - Q_{(a,b)}^{\text{old}}(x_i, x_j)]^2
   \]
   and repeat the above three steps until the residual is less than some cutoff value (e.g., 10\(^{-9}\));
2. Obtain new estimates of \( \pi \) using the latest converged \( Q \)-matrices;
3. Calculate \( \Delta \log L \) (i.e., the difference in log-likelihood);
4. Repeat Steps 1, 2, and 3 until \( \Delta \log L \) is less than some cutoff value (e.g., 10\(^{-3}\)).

**Algorithm for the SBH + I\(_2\) and RBH + I\(_2\) Models**

Choose initial estimates for the proportion of invariable sites (\( \beta \)) and the nucleotide composition of the invariable sites \( \pi^{\text{inv}} = \{\pi_A^{\text{inv}}, \pi_C^{\text{inv}}, \pi_G^{\text{inv}}, \pi_T^{\text{inv}}\} \). The choice of initial values for these parameters is not crucial but we typically set \( \beta \) to a quarter of the proportion of constant sites and use a uniform nucleotide composition for the invariable sites.

1. Remove a subset of the constant sites consistent with the latest estimates of \( \beta \) and \( \pi^{\text{inv}} \) to obtain \( N_{\text{var}} \). In other words, \( N_{\text{var}} = N - \sum_{x_i \in B} N\beta\pi_x^{\text{inv}} \) with the relevant number of As, Cs, Gs, and Ts removed. We also set the initial estimate of the Lagrange multipliers to be half the number of potentially variable sites;
2. Obtain the converged estimates of the \( Q \)-matrices for only the \( N_{\text{var}} \) sites using the algorithm described earlier;
3. Obtain new estimates of \( \beta \) and \( \pi^{\text{inv}} \);
4. Calculate the change in the total log-likelihood using Equation (24);
5. Repeat Steps 1, 2, 3, and 4 until the change in total log-likelihood is less than some cutoff value (e.g., 10\(^{-3}\)).

**APPENDIX 2**

**EM Approach**

Because we do not know the invariable sites, we introduce a random variable \( Z \) to denote the site-type. Next, we use an EM approach (Dempster et al. 1977) to obtain the expected number of invariable sites and the maximum-likelihood estimates of the parameters for the potentially variable sites.

We introduce notation similar to that used in Chapter 3 of Kung et al. (2004). Let \( Z = \{Z_1, \ldots, Z_N\} \) be the vector of site-type that takes unobservable values \( z = \{z_1, \ldots, z_N\} \) and

\[
z_i = \begin{cases} 
1 & \text{if site } i \text{ is invariable,} \\
0 & \text{otherwise.}
\end{cases}
\]

Let \( Y = \{Y_1, \ldots, Y_N\} \), where \( Y_i \) denotes the random variable corresponding to the pattern at site \( i \) and let \( y \) denotes the observed vector of patterns. Let \( \theta \) denote the vector of all parameters in the SBH + I\(_2\) or RBH + I\(_2\) model, \( f_0(y) = P(Y = y), g_0(z) = P(Z = z), p_0(y, z) = P_0(Y = y, Z = z), \) and \( w_0(z|y) = P_0(Z = z|Y = y) \). Since the sites are assumed to be independent,

\[
\log f_0(y) = \sum_{i=1}^{N} \log f_0(y_i),
\]

where \( N \) is the total number of sites. Since \( p_0(y, z) = f_0(y)w_0(z|y) \),

\[
\log f_0(y) = \log p_0(y, z) - \log w_0(z|y).
\]

Replacing \( (y, z) \) by random variables \( (Y, Z) \) and taking the conditional expectation with respect to \( Y = y \), we get

\[
\log f_0(y) = E_{\theta'}[\log p_0(Y, Z)|Y = y] - E_{\theta'}[\log w_0(Z|Y)|Y = y] = S(\theta|\theta') - R(\theta, \theta'),
\]

where \( \theta' \) is a specific value of \( \theta \). It should be noted that Equation (A.3) does not depend on the unknown \( z \).

**Calculation of \( S(\theta|\theta') \).**—We define a site that has the same nucleotide in all the sequences as constant and only a fraction of the constant sites are invariable. Under the assumption that all sites are independent and identically distributed, we can reorder \( y \) such that the first \( N_1 \) sites \( y_1, \ldots, y_{N_1} \) are potentially variable and the remaining sites are constant sites. Now

\[
\log p_0(y, z) = \sum_{i=1}^{N_1} \log [g_0(0)v_0(y_i|0)] + \sum_{i=N_1+1}^{N} I(z_i = 0) \log [g_0(0)v_0(y_i|0)] + \sum_{i=N_1+1}^{N} I(z_i = 1) \log [g_0(1)v_0(y_i|1)],
\]

(A.4)
where \( I(p) \) takes the value 1 if the proposition \( p \) is true and 0 otherwise. Therefore,

\[
S(\theta | \theta') = \sum_{i=1}^{N_1} \log [g_\theta(0)v_\theta(y_i|0)] \\
+ \sum_{i=N_1+1}^{N} \mathbf{E}_\theta'[I(Z_i = 0)|Y_i = y_i] \log [g_\theta(0)v_\theta(y_i|0)] \\
+ \sum_{i=N_1+1}^{N} \mathbf{E}_\theta'[I(Z_i = 1)|Y_i = y_i] \log [g_\theta(1)v_\theta(y_i|1)].
\]

(A.5)

Noting that \( \mathbf{E}_\theta'[I(Z=0)|Y=y]=w_{\theta'}(0|y) \) and \( \mathbf{E}_\theta'[I(Z=1)|Y=y]=w_{\theta'}(1|y) \),

\[
S(\theta | \theta') = \sum_{i=1}^{N_1} \log [g_\theta(0)v_\theta(y_i|0)] \\
+ \sum_{i=N_1+1}^{N} w_{\theta'}(0|y_i) \log [g_\theta(0)v_\theta(y_i|0)] \\
+ \sum_{i=N_1+1}^{N} w_{\theta'}(1|y_i) \log [g_\theta(1)v_\theta(y_i|1)].
\]

(A.6)

Calculation of \( R(\theta | \theta') \).—Because the class information is hidden only for those sites that appear unchanged,

\[
\log w_\theta(z|y) = \sum_{i=1}^{N} I(z_i = 0) \log w_\theta(0|y_i) \\
+ \sum_{i=N_1+1}^{N} I(z_i = 1) \log w_\theta(1|y_i)
\]

and

\[
R(\theta | \theta') = \sum_{i=N_1+1}^{N} w_{\theta'}(0|y_i) \log w_\theta(0|y_i) \\
+ \sum_{i=N_1+1}^{N} w_{\theta'}(1|y_i) \log w_\theta(1|y_i).
\]

(A.7)

Estimating \( \theta \).—Let \( K \) denote the number of taxa and \( \mathcal{H} \) denote the set of column vectors of length \( K \), all elements of a vector being identical and equal to \( A, C, G, \) or \( T \). Then,

\[
N_{\text{var}} = N^h w_{\theta'}(0|h)
\]

and \( N_{\text{inv}} = N^h w_{\theta'}(1|h) \),

(A.9)

where \( N^h \) denotes the total number of sites of type \( h \), \( N_{\text{var}} \) and \( N_{\text{inv}} \) denote the expected number of potentially variable and invariant sites, respectively, of type \( h \), and

\[
w_{\theta'}(z|y) = \frac{p_{\theta'}(y,z)}{f_\theta(y)} = \frac{g_{\theta'}(z)v_{\theta'}(y|z)}{g_{\theta'}(0)v_{\theta'}(y|0) + g_{\theta'}(1)v_{\theta'}(y|1)}.
\]

(A.10)

Let \( \theta = (Q, \phi) \), where \( Q \) denotes the vector of all joint probability distribution matrices (\( Q \)-matrices) for the potentially variable sites and \( \phi \) denotes the remaining parameters namely, the probability of a site being invariant and the conditional probability of a site being \( A, C, G, \) or \( T \) given that the site is invariant. The parameters \( Q \) and \( \phi \) are estimated using the maximization step and the expectation step, respectively, of the EM algorithm.

Maximization step.—Consider first the Barry and Hartigan model (Barry and Hartigan 1987; Jayaswal et al. 2005) for estimating \( Q \)-matrices. Their model requires the maximization of the function

\[
\log f_\theta(y) + \lambda \left( 1 - \sum_{x_i \in B} \sum_{x_j \in B} Q_{(a,b)}(x_a, x_b) \right)
\]

with respect to the \( Q \)-matrices by equating the partial derivatives to 0. The Lagrange multiplier \( \lambda \) corresponds to the constraint that all the elements of a \( Q \)-matrix must add to 1. Since \( R(\theta | \theta') \) and the third sum in Equation (A.6) do not depend on \( Q \), the part of Equation (A.11) that is required for estimating the \( Q \)-matrices is

\[
\sum_{i=1}^{N_1} \log [g_\theta(0)v_\theta(y_i|0)] + \sum_{i=N_1+1}^{N} w_{\theta'}(0|y_i) \log [g_\theta(0)v_\theta(y_i|0)] \\
+ \lambda \left( 1 - \sum_{x_i \in B} \sum_{x_j \in B} Q_{(a,b)}(x_a, x_b) \right).
\]

(A.12)

Keeping the terms \( w_{\theta'}(0|y_i) \) and \( N_{\text{var}}^h \) estimated using Equations (A.9) and (A.10) constant during the maximization step, we obtain new estimates of \( Q \)-matrices as follows:

\[
Q_{(a,b)}^{\text{new}}(x_a, x_b) = \frac{Q_{(a,b)}(x_a, x_b)}{\lambda} \left[ \sum_{i=1}^{N_1} v_{\theta}(y_i|0) \right] \\
+ \sum_{i=N_1+1}^{N} w_{\theta'}(0|y_i) \frac{v_{\theta}'(y_i|0)}{v_{\theta}(y_i|0)} \\
= \frac{Q_{(a,b)}(x_a, x_b)}{\lambda} \left[ \sum_{i=1}^{N_1} v_{\theta}(y_i|0) \right] \\
+ \sum_{h \in \mathcal{H}} N_{\text{var}}^h \frac{v_{\theta}(h|0)}{v_{\theta}(h|0)} \\
= \frac{Q_{(a,b)}(x_a, x_b)}{\lambda} \sum_{i=1}^{N_1} v_{\theta}(y_i|0),
\]

(A.13)
where $N_{\text{var}} = N_1 + \sum_{h \in H} N_{\text{var}}^h$ and $v_\theta'(y_i | 0)$ is the first-order partial derivative of $v_\theta(y_i | 0)$ with respect to $Q(a_h, b_h)$. Thus, the Barry and Hartigan algorithm has to be applied only to the potentially variable sites $N_{\text{var}}$ to estimate $Q$-matrices. For the SBH and RBH models, the $Q$-matrices are determined using Equations (10) and (13), respectively. Equation (A.13) is used iteratively to obtain new values for $Q$-matrices till the log-likelihood converges.

**Expectation step.**—Now we will estimate the parameter $\phi$ as follows:

$$f_\theta(h) = g_\theta(0)v_\theta(h|0) + g_\theta(1)v_\theta(h|1)$$  \hspace{1cm} (A.14)

and $f_\theta(h) = E_\theta[N^h / N]$.  \hspace{1cm} (A.15)

Replacing $E_\theta(N^h / N)$ by the observed value $N^h / N$, we obtain

$$N^h / N = g_\theta(0)v_\theta(h|0) + g_\theta(1)v_\theta(h|1).$$  \hspace{1cm} (A.16)

Noting that $\sum_{h \in H} v_\theta(h|1) = 1$ and $v_\theta(h|0)$ is a function of the $Q$-matrices obtained during the maximization step, we can obtain $g_\theta(1)$ and $v_\theta(h|1)$ using Equations (22) and (23). The revised estimates of $\theta$ are substituted in Equation (A.10) to obtain values of $w_\theta'(0|h)$ and the expected number of potentially variable but unchanged sites $N_{\text{var}}^h$. The expectation step and maximization step are repeated till $\log f_\theta(y)$ converges. For a theoretical proof of convergence under the EM approach, refer to Dempster et al. (1977).