Estimation of Phylogeny and Invariant Sites under the General Markov Model of Nucleotide Sequence Evolution

Vivek Jayaswal1,2,3 John Robinson2,3 and Lars Jermiin1,2,4

1Sydney Bioinformatics, University of Sydney, NSW 2006, Australia
2Centre for Mathematical Biology, University of Sydney, NSW 2006, Australia
3School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
4School of Biological Sciences, University of Sydney, NSW 2006, Australia; E-mail: lars.jermiin@usyd.edu.au

Abstract.—The models of nucleotide substitution used by most maximum likelihood-based methods assume that the evolutionary process is stationary, reversible, and homogeneous. We present an extension of the Barry and Hartigan model, which can be used to estimate parameters by maximum likelihood (ML) when the data contain invariant sites and there are violations of the assumptions of stationarity, reversibility, and homogeneity. Unlike most ML methods for estimating invariant sites, we estimate the nucleotide composition of invariant sites separately from that of variable sites. We analyze a bacterial data set where problems due to lack of stationarity and homogeneity have been previously well noted and use the parametric bootstrap to show that the data are consistent with our general Markov model. We also show that estimates of invariant sites obtained using our method are fairly accurate when applied to data simulated under the general Markov model.

In [variant sites; maximum likelihood; nonhomogeneous process; nonstationary process; nucleotide substitution; phylogenetics.]

Many maximum-likelihood-based methods for phylogenetic inference rely on explicit models of nucleotide substitution, the common underlying assumption being that the process is Markovian (i.e., the conditional probability of change at a given site depends only on the current state and is independent of the previous states). In addition, if we assume that the Markovian process is stationary, homogeneous, and reversible (for details, see Jayaswal et al., 2005; Ababneh et al., 2006b) and that each site furthermore is independent and identically distributed (iid), then we obtain the family of time-reversible substitution models, which ranges from the simple model of Jukes and Cantor (1969) to the general time-reversible (GTR) model of Lanave et al. (1984).

One of the factors known to bias the phylogenetic inference is compositional heterogeneity among sequences (e.g., Lockhart et al., 1994; Foster and Hickey, 1999; Chang and Campbell, 2000; Conant and Lewis, 2001; Jermiin et al., 2004), which implies a violation of the assumption of stationarity, and so also a violation of the assumptions of homogeneity and reversibility. Even in the absence of compositional heterogeneity, rate heterogeneity among lineages can confound phylogenetic estimates (e.g., Felsenstein, 1978; Hasegawa et al., 1991; Steel et al., 1993; Ho and Jermiin, 2004), so there is a need for more general nucleotide substitution models that relax the assumptions of stationarity, reversibility, and homogeneity. Based on the maximum-likelihood criterion, such models have been proposed by Barry and Hartigan (1987), Yang and Roberts (1995), Galtier and Gouy (1998), Galtier et al. (1999), and Foster (2004).

Yang and Roberts (1995) considered separate HKY models (Hasegawa et al., 1985) along different edges of the tree by varying the frequency parameters but encountered a convergence problem even for as few as four taxa. Based on Tamura's (1992) model, which is simpler than the HKY model, Galtier and Gouy (1998) developed a model that can account for nonstationary and nonhomogeneous processes and used it to infer the phylogeny of 16 bacterial ribosomal RNA sequences. Their model was generalized by Galtier et al. (1999) to account for rate heterogeneity across sites. Though none of these models require global homogeneity, they still assume local homogeneity; i.e., the evolutionary process over a particular edge of the tree cannot change with time—this is also true for Foster's (2004) model. All of these models can therefore be considered special cases of Barry and Hartigan's (1987) model, which does not require stationarity, reversibility, and homogeneity (global or local).

Apart from relaxing the assumptions of stationarity, reversibility, and homogeneity, the Markovian models of molecular evolution can be made more realistic by allowing for different rates of evolution at different sites along a sequence. Ignoring this information can result in an underestimation of edge lengths and biases in the estimation of other parameters (e.g., Wakeley, 1994; Yang et al., 1994). Two of the most commonly used approaches for considering rate heterogeneity across sites are (i) grouping of sites into variable and fixed sites, with the processes at variable sites being $iid$ (e.g., Lockhart et al., 1996; Steel et al., 2000; Swofford, 2002; Guindon and Gascuel, 2003); and (ii) using a continuous or discretized $\Gamma$-distribution to integrate over rates at individual sites (e.g., Yang, 1993, 1994). Although the family of time-reversible substitution models has been extended to incorporate both these approaches, and Yang and Roberts' (1995) models have been extended to incorporate the latter approach, the Barry and Hartigan (1987) model, henceforth called the BH model, currently assumes that all sites are $iid$.

Jayaswal et al. (2005) implemented the BH model and highlighted the advantages of using this model as opposed to the family of time-reversible substitution models, such as (i) obtaining a good fit even when the process is not stationary, reversible, and homogeneous; (ii) checking the assumption of reversibility; and (iii) finding instances of convergent evolution more frequently compared to the family of time-reversible
substitution models. Although the BH model is promising for the above three reasons, it is of limited use as it assumes that the processes at the sites are iid. This assumption can be relaxed by assuming that the nucleotides at some sites remain constant; i.e., invariant over time. In this paper, we use the term “unchanged” to denote those sites that have the same nucleotide in all sequences, and the term “invariant” to denote unchanged sites that are not allowed to change.

The proportion of invariant sites can be estimated using a capture-recapture method (Sidow et al., 1992; Steel et al., 2000) or a maximum-likelihood-based approach. Phylogenetic software packages, such as SplitsTree (Huson and Bryant, 2006), use the former method, whereas other packages, such as PAUP* (Swofford, 2002) and PHYML (Guindon and Gascuel, 2003), use the latter approach to estimate the proportion of sites that are invariant. The capture-recapture methods estimate the proportion of sites that are invariant but not their nucleotide composition. Therefore, further analysis of the data, namely the estimation of tree topology and edge lengths, which requires removal of the invariant sites, is based on the assumption that the proportion of invariant As, Cs, Gs, and Ts is (i) equal to 0.25; (ii) equal to the proportion of As, Cs, Gs, and Ts in the data set; or (iii) equal to the proportion of As, Cs, Gs, and Ts in the sites that appear unchanged. As the correct estimate of the nucleotide composition of invariant sites may differ from these proportions, it is desirable to estimate the parameters for the variable and invariant sites separately, as pointed out by Waddell and Steel (1997). Unlike the capture-recapture methods, the maximum-likelihood methods can be used directly to estimate the nucleotide composition of invariant sites.

In this paper, we describe an approach for estimating invariant sites without the assumptions of stationarity, reversibility, and homogeneity. In doing so, we assume that (i) the sites in an alignment can be divided into two groups—variable and invariant (all sites excluding the invariant sites are said to be variable); (ii) all variable sites evolve under independent and identically distributed Markov processes. Next, we analyze an alignment of bacterial ribosomal RNA sequences using a version of the BH model that accounts for the presence of invariant sites (i.e., the BH + I model). Finally, parametric bootstrapping is used to show (i) that the bacterial data set is consistent with BH + I model; and (ii) that the maximum-likelihood estimates of invariant sites are close to the actual values in data generated under the model.

METHOD

We introduce notation similar to that used by Jayaswal et al. (2005). Let $T$ denote an unrooted binary tree with $l$ leaves, $l - 2$ internal nodes, and $2l - 3$ edges, for $l \geq 2$. Denote leaves by $L = \{-1, \ldots, -l\}$ and internal nodes by $I = \{1, \ldots, l - 2\}$. Denote the set of all nodes $V = L \cup I$ and the set of all edges by $E = \{(a, b): a, b \in V$ and adjacent}. 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 $B = \{A, C, G, T\}$ and $X_a$ and $X_b$ be the nucleotides at nodes $a$ and $b$ of the edge $(a, b)$. The joint probability along the edge $(a, b)$ is

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

and the marginal probability of $x_a$ at node $a$ is

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

Let $L_{(a, b)}$ and $T_{(a, b)} \equiv I \cap T_{(a, b)}$ denote the sets of leaf nodes and internal nodes, respectively, in $T_{(a, b)}$, then $x_{L_{(a, b)}}$ and $x_{T_{(a, b)}}$ denote the vectors of the values of nucleotides in $L_{(a, b)}$ and $T_{(a, b)}$, respectively, and

$$Q_{L_{(a, b)}}(x_{L_{(a, b)}}) = \sum_{x_{T_{(a, b)}}} Q_{T_{(a, b)}}(x_{T_{(a, b)}}). \quad (3)$$

Take the model to be Markovian, so that, given $(X_a, X_b) = (x_a, x_b)$, the conditional distribution of the nucleotides on the leaves of the rooted subtrees $T_{(a, b)}$ and $T_{(b, a)}$, obtained by deleting edge $(a, b)$, are independent. At each site $i = 1, \ldots, N$ the value of a nucleotide at the $r$th leaf, $x_r$, is known, but at internal nodes the nucleotide can take any value in $B$. Let $B_r = \{x_r\}$ if $r \in L$, and $B$ if $r \in I$. Then the joint probability distribution of leaf nodes at site $i$ can be expressed as

$$L_i = Q_{L_i}(x_{L_i}) = \sum_{x_a \in B} \sum_{x_b \in B} I(x_a \in B_{ai}, x_b \in B_{bi}) Q_{(a, b)}(x_a, x_b) \times (x_a, x_b) P(L_{(a, b)}|x_a) P(L_{(b, a)}|x_b) \quad (4)$$

where $I(x_a \in B_{ai}, x_b \in B_{bi})$ is an indicator function that takes the value 1 if both conditions are satisfied, 0 otherwise, and

$$P(L_{(a, b)}|x_a) = \frac{Q_{L_{(a, b)}}(x_{L_{(a, b)}}; x_a)}{Q_a(x_a)}. \quad (5)$$

Note that (4) and (5) hold for all edges $(a, b) \in E$. Therefore, it is necessary that

$$Q_a(x_a) = \sum_{x_b \in B} Q_{(a, b)}(x_a, x_b) = \sum_{x_b \in B} Q_{(a, c)}(x_a, x_b) \quad (6)$$

where $b$, $c$, and $d$ are the three nodes adjacent to the internal node $a$.

Differentiating Equation (4) with respect to $Q_{(a, b)}(x_a, x_b)$, we obtain

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

So far we have concerned ourselves with the process at a single site. We will now focus on the processes at
different sites and use the following notation to describe site-specific information:

(i) Let \( Y = m \times N \) denote the matrix of aligned nucleotide sequences with \( m \) being the number of sequences and \( N \) being the number of sites in the alignment. Then, \( y_i \) denotes the \( i \)th column of the data matrix.

(ii) Let \( N_2 \) denote the number of sites that appear unchanged, i.e., sites having the same nucleotide in all sequences, and let \( N_1 = N - N_2 \) denote the remaining number of sites. Because the sites are assumed to be independent, we can rearrange them without loss of generality such that the last \( N_2 \) sites correspond to the unchanged sites.

(iii) Let \( \beta \) denote the probability of a site being invariant and \( \alpha \) denote the probability that a site is variable.

(iv) Let \( \mathcal{H} \) denote the set of column vectors of length \( m \), all elements of a vector being identical and equal to A, C, G, or T.

(v) Let \( P(y_i|\text{inv}) \) denote the probability that the \( i \)th site is of type \( h \in \mathcal{H} \), given that it is an invariant site. Then, \( P(y_i|\text{inv}) = P(h|\text{inv}) \) for all sites having the pattern \( y_i = h \).

Now, the overall log-likelihood can be written as

\[
\log L = \sum_{i=1}^{N} \log L_i^* \tag{7}
\]

where

\[
L_i^* = \alpha P(y_i|\text{var}) + \beta P(y_i|\text{inv}) \tag{8}
\]

Because in the current model it is assumed that an invariant site does not change over the entire phylogenetic tree, \( P(y_i|\text{inv}) = 0 \) when \( y_i \not\in \mathcal{H} \). Therefore, (8) can be rewritten as

\[
L_i^* = \begin{cases} 
\alpha P(y_i|\text{var}) + \beta P(y_i|\text{inv}) & \text{if } y_i \in \mathcal{H} \\
\alpha P(y_i|\text{var}) & \text{if } y_i \not\in \mathcal{H}
\end{cases} \tag{9}
\]

For a site that is not invariant, \( P(y_i|\text{var}) = L_i \) and (7) can be rewritten as

\[
\log L = \sum_{i=1}^{N_1} \log(\alpha L_i) + \sum_{i=N_1+1}^{N} \log(\alpha L_i + \beta P(y_i|\text{inv})) \tag{9}
\]

where \( N_1 \) is a data-dependent variable.

Because the maximum-likelihood estimates of the parameters have to be determined subject to the constraints that \( \sum_{x_a \in A} \sum_{x_b \in B} Q_{(a,b)}(x_a, x_b) = 1, \alpha + \beta = 1 \) and \( \sum_{h \in \mathcal{H}} P(h|\text{inv}) = 1 \), we introduce Lagrange multipliers for each of the three constraints and maximize

\[
\log L + \lambda_1 \left( 1 - \sum_{x_a \in A} \sum_{x_b \in B} Q_{(a,b)}(x_a, x_b) \right) + \lambda_2 (1 - \alpha - \beta) + \lambda_3 \left( 1 - \sum_{h \in \mathcal{H}} P(h|\text{inv}) \right). \tag{10}
\]

Differentiating Equation (10) with respect to \( Q_{(a,b)}(x_a, x_b) \) and setting the derivative equal to 0, we obtain

\[
\sum_{i=1}^{N_1} \frac{L_i(x_a, x_b)}{L_i} + \sum_{i=N_1+1}^{N} \frac{\alpha L_i(x_a, x_b)}{\alpha L_i + \beta P(y_i|\text{inv})} = \lambda_1. \tag{11}
\]

Using Equation (11), we can obtain a new value of \( Q_{(a,b)}(x_a, x_b) \) as follows

\[
Q_{(a,b)}^\text{new}(x_a, x_b) = \frac{Q_{(a,b)}(x_a, x_b)}{\lambda_1} \times \left[ \sum_{i=1}^{N_1} \frac{L_i(x_a, x_b)}{L_i} + \sum_{i=N_1+1}^{N} \frac{\alpha L_i(x_a, x_b)}{\alpha L_i + \beta P(y_i|\text{inv})} \right]. \tag{12}
\]

As in Section 3.2 of Jayaswal et al. (2005), it can be shown that (12) satisfies the constraint of internal consistency, i.e., (6) holds for all internal nodes.

Now,

\[
\beta^\text{new} = \frac{\beta}{\lambda_2} \sum_{i=N_1+1}^{N} \frac{P(y_i|\text{inv})}{\alpha L_i + \beta P(y_i|\text{inv})} \tag{13}
\]

\[
\alpha^\text{new} = \frac{\alpha}{\lambda_2} \left[ \frac{N_1}{\alpha} + \sum_{i=N_1+1}^{N} \frac{L_i}{\alpha L_i + \beta P(y_i|\text{inv})} \right] \tag{14}
\]

and

\[
P^\text{new}(h|\text{inv}) = \frac{P(h|\text{inv})}{\lambda_3} \sum_{i=N_1+1}^{N} \frac{I(y_i = h)\beta}{\alpha L_i + \beta P(y_i|\text{inv})} \tag{15}
\]

where \( I(y_i = h) \) is an indicator function that takes the value 1 if \( y_i = h \), and 0 otherwise.

Using the appropriate constraint from Equation (10), the values of \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) can be determined. For example,

\[
\alpha^\text{new} + \beta^\text{new} = 1.
\]

Therefore, using Equations (13) and (14),

\[
\frac{\beta}{\lambda_2} \sum_{i=N_1+1}^{N} \frac{P(y_i|\text{inv})}{\alpha L_i + \beta P(y_i|\text{inv})} + \frac{\alpha}{\lambda_2} \left[ \frac{N_1}{\alpha} + \sum_{i=N_1+1}^{N} \frac{L_i}{\alpha L_i + \beta P(y_i|\text{inv})} \right] = 1
\]
or
\[
\lambda_2 = N_1 + \sum_{i=N_1+1}^N \frac{\alpha L_i}{\alpha L_i + \beta P(y_i|\text{inv})} + \sum_{i=N_1+1}^N \frac{\beta P(y_i|\text{inv})}{\alpha L_i + \beta P(y_i|\text{inv})}.
\] (16)

Thus, using Equations (13), (14), and (16), new values of \(\alpha\) and \(\beta\) can be obtained.

Note that the estimates obtained at the end of each iteration using Equations (12) to (15) are always non-negative because all the terms on the right-hand side of these equations, including the Lagrange multipliers, are non-negative.

The initial values for the \(Q\)-matrices are chosen such that the diagonal elements are greater than the off-diagonal elements because, as noted by Jayaswal et al. (2005), such \(Q\)-matrices tend to converge to a common local maximum. If extreme values of \(Q\)-matrices are chosen with the off-diagonal elements being greater than the diagonal ones, then we can obtain convergence to local maxima that differ from the most common one. During simulations we observed that the local maxima obtained using extreme values of \(Q\)-matrices were lower than the common maximum. Also, the initial \(Q\)-matrices should satisfy the constraint of internal consistency; i.e., (6) must be satisfied for all internal nodes.

The initial values of the remaining parameters are chosen such that \(0 < \beta, P(A|\text{inv}), P(C|\text{inv}), P(G|\text{inv}), P(T|\text{inv}) < 1\) and \(\sum_{x \in B} P(X|\text{inv}) = 1\). Next, Equations (12) to (15) are used to update the parameter values until convergence is achieved.

In the BH+I model described above, all the parameters are updated simultaneously at the end of each iteration. However, the BH+I model can also be implemented as follows: (i) remove a proportion of the unchanged sites (which are estimated to be invariant) from the data set and use the BH model over the remaining sites to estimate the \(Q\)-matrices; (ii) use the revised \(Q\)-matrices to estimate \(\beta, P(A|\text{inv}), P(C|\text{inv}), P(G|\text{inv}), P(T|\text{inv})\). The two steps are repeated until convergence is achieved. The expectation-maximization (EM) (Dempster at al., 1977; Kung et al., 2004) approach to parameter optimization of the BH+I model essentially reduces to the two steps described in this paragraph. It gives the same result as that obtained using the direct optimization of parameters but the current implementation of the EM algorithm is computationally slower.


Reeves (1992) used an EM algorithm for estimating the parameters for amino acid sequences. However, his programs are not available online and do not appear to have been incorporated into any of the existing phylogenetic software packages.

**RESULTS**

**Bacterial Data**

We analyzed a bacterial data set, previously examined by Jayaswal et al. (2005) and Ababneh et al. (2006a), using the BH+I model. The data set comprises an alignment of 1238 nucleotides from the 16S ribosomal RNA genes (with abbreviated names given in parentheses) of: *Aquifex pyrophilus* (Apyr), *Thermotoga maritima* (Tmar), *Thermus thermophilus* (Tthe), *Deinococcus radiodurans* (Drad), and *Bacillus subtilis* (Bsub). Using methods described in Ababneh et al. (2006a), these data were found to violate the assumptions of stationarity, reversibility, and homogeneity (Jayaswal at al., 2005).

Table 1 shows some of the phylogenetic results obtained using the BH and BH+I models. The number of parameters for the BH model are 241—33, where \(l\) denotes the number of leaf nodes (Jayaswal at al., 2005), whereas the BH+I model has an additional four parameters. The large increase in log-likelihood values (\(>93\)) for an additional four parameters provides evidence that the bacterial data set has a proportion of invariant sites that were not properly accounted for in previous phylogenetic analyses (Galtier and Gouy, 1995; Jayaswal et al., 2005; Ababneh et al., 2006a).

The ranks of the tree topologies under the two models differ slightly—e.g., \(T_7\) and \(T_8\) have higher log-likelihood values compared to \(T_3\) under the BH model but not under the BH+I model. Also, the estimate of \(\beta\) varies only slightly with the tree topology. This is in agreement with the observation that although maximum-likelihood estimates of invariant sites are tree dependent, large deviations are unlikely (Sullivan at al., 1996; Lockhart et al., 1998). The most likely tree obtained using the BH+I model is the same as that inferred by Galtier and Gouy (1995), Gupta (2000), Foster (2004), Jayaswal et al. (2005), and Ababneh et al. (2006a).

Table 2 shows the log-likelihoods obtained under the GTR, GTR+I, and GTR+Γ models for the 15 unrooted

<table>
<thead>
<tr>
<th>Number</th>
<th>Topology</th>
<th>BH</th>
<th>BH+I</th>
<th>(\beta^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T_1)</td>
<td>(Bsub,(Tmar,Ap)ry),(Tthe,Drad))</td>
<td>-4289.511</td>
<td>-4193.841</td>
<td>0.554</td>
</tr>
<tr>
<td>(T_2)</td>
<td>(Tthe,(Tmar,Ap)ry),(Drad,Bsub))</td>
<td>-4296.220</td>
<td>-4203.027</td>
<td>0.553</td>
</tr>
<tr>
<td>(T_3)</td>
<td>(Tmar,(Bsub,Ap)ry),(Tthe,Drad))</td>
<td>-4312.360</td>
<td>-4206.990</td>
<td>0.562</td>
</tr>
<tr>
<td>(T_4)</td>
<td>(Ap)ry,(Tmar,Bsub),(Drad,Tthe)</td>
<td>-4317.686</td>
<td>-4207.546</td>
<td>0.564</td>
</tr>
<tr>
<td>(T_5)</td>
<td>(Tmar,(Bsub,Ap)ry),(Drad,Tthe)</td>
<td>-4330.780</td>
<td>-4224.162</td>
<td>0.566</td>
</tr>
<tr>
<td>(T_6)</td>
<td>(Ap)ry,(Bsub,Drad),(Tthe,Tmar)</td>
<td>-4331.682</td>
<td>-4224.738</td>
<td>0.564</td>
</tr>
<tr>
<td>(T_7)</td>
<td>(Drad,(Tmar,Ap)ry),(Tthe,Bsub))</td>
<td>-4356.860</td>
<td>-4217.596</td>
<td>0.569</td>
</tr>
<tr>
<td>(T_8)</td>
<td>(Drad,(Tmar,Bsub),(Tthe,Ap))</td>
<td>-4360.567</td>
<td>-4231.240</td>
<td>0.577</td>
</tr>
<tr>
<td>(T_9)</td>
<td>(Drad,(Tmar,Ap)ry),(Tthe,Bsub))</td>
<td>-4361.190</td>
<td>-4235.681</td>
<td>0.576</td>
</tr>
<tr>
<td>(T_{10})</td>
<td>(Tthe,(Bsub,Ap)ry),(Tmar,Bsub))</td>
<td>-4368.976</td>
<td>-4232.252</td>
<td>0.579</td>
</tr>
<tr>
<td>(T_{11})</td>
<td>(Tthe,(Bsub,Drad),(Tmar,Ap)ry)</td>
<td>-4373.788</td>
<td>-4237.007</td>
<td>0.580</td>
</tr>
<tr>
<td>(T_{12})</td>
<td>(Bsub,(Tmar,Ap)ry),(Tthe,Drad))</td>
<td>-4377.168</td>
<td>-4240.701</td>
<td>0.582</td>
</tr>
<tr>
<td>(T_{13})</td>
<td>(Bsub,(Tmar,Drad),(Tthe,Ap))</td>
<td>-4378.561</td>
<td>-4240.177</td>
<td>0.583</td>
</tr>
<tr>
<td>(T_{14})</td>
<td>(Ap)ry,(Tmar,Bsub),(Drad,Tthe)</td>
<td>-4385.884</td>
<td>-4240.816</td>
<td>0.585</td>
</tr>
<tr>
<td>(T_{15})</td>
<td>(Tmar,(Bsub,Ap)ry),(Drad,Ap))</td>
<td>-4387.239</td>
<td>-4240.960</td>
<td>0.586</td>
</tr>
</tbody>
</table>

*Probability of a site being invariant.*

*Barry and Hartigan model.*

*Barry and Hartigan model with invariant sites.*
Table 2. Log-likelihood values for the 15 unrooted binary trees under the GTR*, GTR+I*, and GTR+I+* models.

<table>
<thead>
<tr>
<th>Number</th>
<th>Topology</th>
<th>GTR*</th>
<th>GTR+I*</th>
<th>GTR+I+*</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>(Bsub,(Tmar,Apyr),(Tthe,Drad))</td>
<td>-4414.97</td>
<td>-4321.17</td>
<td>-4317.28</td>
</tr>
<tr>
<td>T2</td>
<td>(Tthe,(Tmar,Apyr),(Drad,Bsub))</td>
<td>-4401.42</td>
<td>-4316.18</td>
<td>-4313.64</td>
</tr>
<tr>
<td>T3</td>
<td>(Tmar,(Bsub,Apyr),(Tthe,Drad))</td>
<td>-4442.74</td>
<td>-4333.79</td>
<td>-4330.65</td>
</tr>
<tr>
<td>T4</td>
<td>(Apyr,(Tmar,Bsub),(Drad,Tthe))</td>
<td>-4446.48</td>
<td>-4333.94</td>
<td>-4330.90</td>
</tr>
<tr>
<td>T5</td>
<td>(Tmar,(Bsub,Drad),(Tthe,Apyr))</td>
<td>-4345.76</td>
<td>-4333.14</td>
<td>-4331.51</td>
</tr>
<tr>
<td>T6</td>
<td>(Apyr,(Bsub,Drad),(Tthe,Tmar))</td>
<td>-4437.26</td>
<td>-4333.86</td>
<td>-4331.55</td>
</tr>
<tr>
<td>T7</td>
<td>(Drad,(Tmar,Apyr),(Tthe,Bsub))</td>
<td>-4433.66</td>
<td>-4333.86</td>
<td>-4331.55</td>
</tr>
<tr>
<td>T8</td>
<td>(Drad,(Tmar,Bsub),(Tthe,Apyr))</td>
<td>-4443.43</td>
<td>-4333.66</td>
<td>-4331.55</td>
</tr>
<tr>
<td>T9</td>
<td>(Drad,(Tmar,Apyr),(Bsub,Tmar))</td>
<td>-4446.37</td>
<td>-4333.79</td>
<td>-4331.55</td>
</tr>
<tr>
<td>T10</td>
<td>(Tmar,(Bsub,Drad),(Tthe,Apyr))</td>
<td>-4401.42</td>
<td>-4316.18</td>
<td>-4313.64</td>
</tr>
<tr>
<td>T11</td>
<td>(Bsub,(Tmar,Tthe),(Drad,Apyr))</td>
<td>-4488.38</td>
<td>-4355.45</td>
<td>-4353.02</td>
</tr>
<tr>
<td>T12</td>
<td>(Bsub,(Tmar,Tthe),(Apyr,Drad))</td>
<td>-4489.94</td>
<td>-4356.66</td>
<td>-4354.18</td>
</tr>
<tr>
<td>T13</td>
<td>(Apyr,(Tmar,Tthe),(Drad,Bsub))</td>
<td>-4500.52</td>
<td>-4357.13</td>
<td>-4354.41</td>
</tr>
<tr>
<td>T14</td>
<td>(Tmar,(Bsub,Apyr),(Tthe,Drad))</td>
<td>-4437.26</td>
<td>-4333.14</td>
<td>-4331.51</td>
</tr>
<tr>
<td>T15</td>
<td>(Tthe,(Drad,Apyr),(Bsub,Tmar))</td>
<td>-4490.74</td>
<td>-4357.13</td>
<td>-4354.41</td>
</tr>
</tbody>
</table>

*General time-reversible model.
+General time-reversible model with invariant sites.
++General time-reversible model with discretized i-distribution of rates.

Tree topologies. The ranks of the trees vary slightly from those obtained under the BH or BH+I model. For example, unlike the BH and BH+I models, T2 is the most likely tree under the GTR models.

To test whether a general Markov model with invariant sites is sufficient to explain the evolutionary process, we used a modification of Cox’s (1962) test as described by Goldman (1993) for nested models. Using the maximum-likelihood estimates of the parameters (i.e., Q-matrices, β, P(A|inv), P(C|inv), P(G|inv), P(T|inv)) obtained for T1, we generated 1,000 pseudo-data sets. For each pseudo-data set, we calculated δ = log L_U - log L_I, where log L_U denotes the unconstrained log-likelihood and log L_I denotes the log-likelihood for T1. The unconstrained log-likelihood log L_U equals ∑_pattern N_pattern log(N_pattern/N), where N_pattern denotes the number of occurrences of a particular pattern in the data set and N denotes the total number of sites.

For the bacterial data set, δ_obs (i.e., the difference between unconstrained log-likelihood and log-likelihood of T1) is 229.93 and the proportion of pseudo-data sets with δ ≥ δ_obs was found to be 0.111 (Fig. 1). Because δ_obs lies within the expected distribution of δ (i.e., the one generated using parametric bootstrap), we have evidence that the bacterial data could have arisen under a general Markov model with invariant sites.

To test the null hypothesis that the GTR+I model is sufficient to explain the bacterial data set against the alternate hypothesis that the BH+I model explains the data set, we generated 1,000 pseudo-data sets under the GTR+I model using Seq-Gen (Rambaut and Grassly 1997) for the parameters corresponding to T2. The pseudo-data sets were analyzed under both the models for T2 and the difference in log-likelihoods varied between 23.85 and 61.01; none of the values were larger than the observed difference of 113.91 (i.e., the difference between -4203.027 and -4316.18; values obtained from Tables 1 and 2). Thus, we have strong evidence against the null hypothesis, implying that the BH+I model provides a better fit for the bacterial data set and the large difference in log-likelihood values between the two models is not due to overparameterization.

To test the null hypothesis that a particular tree, T, is the true tree versus the alternate hypothesis that a tree other than T is the true tree, we applied the SOWH test (Swofford et al., 1996; Goldman et al., 2000; Shi et al., 2005) to each of the 15 unrooted trees. It is a parametric bootstrap test used for hypothesis testing when T is not selected a priori as the true tree, and the model of nucleotide substitution is correct. The test can be performed as follows:

1. generate 1,000 pseudo-data sets under the null hypothesis;
2. calculate δ = log L_ML - log L_T for each pseudo-data set, where log L_ML is the log-likelihood of the most likely tree and log L_T is the log-likelihood of T;
3. calculate p, the proportion of times that δ ≥ δ_obs, where δ_obs is the difference in the log-likelihood values of the most likely tree and T for the original data set. A large p-value would support the null hypothesis; otherwise, there is evidence against it.

The above-mentioned three steps were repeated for each of the 15 unrooted trees. Because T1 is the most likely tree, δ_obs = 0 and the p-value for the null hypothesis—that T1 is the true tree—is 1. The extremely low p-values (Table 3) for the remaining trees provide strong evidence against any tree other than T1 being the true tree.

In addition to the SOWH test, for each tree T_i, i = 1, ..., 15, we also determined the percentage of the pseudo-data sets where the most likely tree and the true tree were the same. Some of the trees were found to have low percentages (Table 3), suggesting that if one of these trees were correct, then the internal edge lengths would be very short (or almost 0). The Q-matrix along an edge
of length 0 would be diagonal, and an analysis of the Q-matrices (estimated on the bacterial data set) for the trees with low percentages showed that this was indeed the case.

Figure 2 shows the labeled nodes for $T_1$ and Table 4 shows the marginal probabilities at these nodes. The marginal probabilities were obtained after removing the invariant sites and can be divided into two groups. The first group comprises the internal nodes (1, 2, and 3) and the leaf nodes *Aquifex*, *Thermotoga*, and *Thermus*. The other group consists of the leaf nodes *Bacillus* and *Deinococcus*. The marginal probabilities of nodes within each group are similar but different from those in the other group. This is even more apparent when the GC content is considered (Fig. 2). The distribution of the marginal probabilities of the nodes belonging to the first group suggests that the subtree excluding the leaf nodes *Bacillus* and *Deinococcus* has a stationary Markov process acting on it. Furthermore, the Q-matrices along the edges of the subtree were found to be symmetric, which suggests that the subtree could be represented by a time-reversible model ($R_2$). The Q-matrices along the edges leading to the leaf nodes *Bacillus* and *Deinococcus* were found to be similar with respect to one another. Table 5 shows the inferred divergence matrices along the two edges (obtained by multiplying the Q-matrices with the number of variable sites). The two matrices have a noticeable substitution bias of G $\rightarrow$ A, C $\rightarrow$ T, and G $\rightarrow$ T over A $\rightarrow$ G, T $\rightarrow$ C, and T $\rightarrow$ G, unlike the remaining edges, which had symmetric Q-matrices. This, coupled with the observation that the internal nodes 2 and 3 have similar marginal probabilities, suggests that another Markov model, $R_2$ (where $R_1 \neq R_2$), could be used to explain the two independent processes along the edges leading to *Bacillus* and *Deinococcus*. Such a scenario, based on two different rate matrices, was proposed by Foster (2004) and found to have a good fit for the bacterial data. Thus, in addition to providing a good fit, the BH+I model can help in identifying simpler models that are consistent with the data.

**Comparison of Models for Estimating Invariant Sites**

The estimate of $\beta$ obtained using the BH+I model was compared with those obtained using other programs that can estimate $\beta$. We used the 1,000 pseudo-data sets generated on $T_1$ and the maximum-likelihood estimates of the parameters (i.e., Q-matrices, $\beta$, $P(A|\text{inv})$, $P(C|\text{inv})$, $P(G|\text{inv})$, $P(T|\text{inv})$) obtained from the bacterial data set and compared the $\beta$ values obtained using the BH+I model with those obtained using the methods implemented in PHYML and SplitsTree. For PHYML, the $\beta$ values were estimated using the GTR+I model and found to be close to those obtained using the BH+I model. In contrast, the capture-recapture method of SplitsTree returned significantly lower estimates of $\beta$ (Table 6).

---

**Table 3. Swofford-Olsen-Wadell-Hillis test results for all 15 unrooted binary trees.**

<table>
<thead>
<tr>
<th>Tree</th>
<th>p-value</th>
<th>PMLTa</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>1.000</td>
<td>99.5%</td>
</tr>
<tr>
<td>$T_2$</td>
<td>0.001</td>
<td>99.8%</td>
</tr>
<tr>
<td>$T_3$</td>
<td>0.005</td>
<td>83.5%</td>
</tr>
<tr>
<td>$T_4$</td>
<td>0.010</td>
<td>70.4%</td>
</tr>
<tr>
<td>$T_5$</td>
<td>0.004</td>
<td>69.8%</td>
</tr>
<tr>
<td>$T_6$</td>
<td>0.003</td>
<td>58.8%</td>
</tr>
<tr>
<td>$T_7$</td>
<td>0.002</td>
<td>32.0%</td>
</tr>
<tr>
<td>$T_8$</td>
<td>0.003</td>
<td>64.6%</td>
</tr>
<tr>
<td>$T_9$</td>
<td>0.001</td>
<td>80.7%</td>
</tr>
<tr>
<td>$T_{10}$</td>
<td>0.001</td>
<td>33.7%</td>
</tr>
<tr>
<td>$T_{11}$</td>
<td>0.003</td>
<td>33.0%</td>
</tr>
<tr>
<td>$T_{12}$</td>
<td>0.004</td>
<td>22.1%</td>
</tr>
<tr>
<td>$T_{13}$</td>
<td>0.001</td>
<td>27.5%</td>
</tr>
<tr>
<td>$T_{14}$</td>
<td>0.002</td>
<td>21.6%</td>
</tr>
<tr>
<td>$T_{15}$</td>
<td>0.002</td>
<td>25.2%</td>
</tr>
</tbody>
</table>

a Swofford-Olsen-Wadell-Hillis test.

**Table 4. Marginal probabilities at the nodes of tree $T_1$.**

<table>
<thead>
<tr>
<th>Node Name</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.143</td>
<td>0.343</td>
<td>0.367</td>
<td>0.147</td>
</tr>
<tr>
<td>2</td>
<td>0.146</td>
<td>0.326</td>
<td>0.373</td>
<td>0.154</td>
</tr>
<tr>
<td>3</td>
<td>0.146</td>
<td>0.324</td>
<td>0.374</td>
<td>0.157</td>
</tr>
<tr>
<td><em>Aquifex</em></td>
<td>0.125</td>
<td>0.367</td>
<td>0.370</td>
<td>0.138</td>
</tr>
<tr>
<td><em>Thermotoga</em></td>
<td>0.108</td>
<td>0.350</td>
<td>0.381</td>
<td>0.160</td>
</tr>
<tr>
<td><em>Thermus</em></td>
<td>0.135</td>
<td>0.347</td>
<td>0.370</td>
<td>0.147</td>
</tr>
<tr>
<td><em>Bacillus</em></td>
<td>0.208</td>
<td>0.258</td>
<td>0.292</td>
<td>0.242</td>
</tr>
<tr>
<td><em>Deinococcus</em></td>
<td>0.206</td>
<td>0.247</td>
<td>0.298</td>
<td>0.249</td>
</tr>
</tbody>
</table>

Percentage of the time most likely tree and the true tree are the same.
The bias may be due to the fact that the sequences were generated under nonstationary and nonhomogeneous conditions and the formula used by this capture-recapture method for estimating invariant sites is not exact when the GTR or more complex substitution models are considered (Steel et al., 2000). For the BH+I model, the actual and estimated $\beta$ values are close (less than one standard deviation apart: see Table 6), but the estimate of $\beta$ is slightly below the expected value.

The invariant sites estimates obtained on $T_1$ using BH+I and GTR+I models were close, but the difference in log-likelihood values was $160.182 \pm 16.286$ (i.e., the mean and sample standard deviation). Although the GTR+I model has 16 degrees of freedom for the five-taxon tree, the BH+I model has 91 degrees of freedom, resulting in 75 additional degrees of freedom for the BH+I model. The large difference in the log-likelihood values cannot be explained only by the increase in the number of degrees of freedom and thus provides additional evidence that the bacterial data have evolved under conditions that are more complex than those accounted for by the GTR+I model. Our result, therefore, corroborates results by Galtier and Gouy (1995), Foster (2004), Jayaswal et al. (2005), and Ababneh et al. (2006a).

The actual and estimated values of the proportion of invariant As, Cs, Gs, and Ts for the pseudo-data sets are shown in Table 7. The estimated values were obtained for $T_1$ and are close to the actual values.

**Table 5.** Substitution biases along two edges of $T_1$ (rows correspond to internal nodes).

<table>
<thead>
<tr>
<th></th>
<th>Bacillus</th>
<th>Deinococcus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>66.7</td>
<td>3.0</td>
</tr>
<tr>
<td>C</td>
<td>11.4</td>
<td>126.0</td>
</tr>
<tr>
<td>G</td>
<td>36.6</td>
<td>7.1</td>
</tr>
<tr>
<td>T</td>
<td>0.0</td>
<td>6.9</td>
</tr>
</tbody>
</table>

**Table 6.** Comparison of $\beta$ estimate obtained using different models for an actual value of 0.554.

<table>
<thead>
<tr>
<th></th>
<th>BH+I*</th>
<th>GTR+I</th>
<th>Capture-recapture*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.543 ± 0.023</td>
<td>0.538 ± 0.021</td>
<td>0.454 ± 0.023</td>
</tr>
</tbody>
</table>

*pValues after ± indicate sample standard deviation.

**Discussion**

The BH model is the most general Markov model of nucleotide evolution under the assumption that the processes at sites are independent and identically distributed ($iid$). As this assumption is often violated by real data, there is a need to consider rate heterogeneity across sites while using the general Markov model. The simplest approach would be to assume that (i) the sites can be divided into two groups depending on whether they are variable or invariant and (ii) variable sites are $iid$. This approach has been described in the current paper.

We used the BH+I model to analyze a set of bacterial ribosomal RNA genes, which have evolved under nonstationary, nonreversible, and nonhomogeneous conditions. Using parametric bootstrap analysis, we found that the optimal tree and the associated parameter values fit the bacterial data well. As the BH+I model provides a good fit, we used the parametric bootstrap to compare tree topologies and showed that there is a strong evidence for $T_1$ being the true tree.

There are several benefits of using the BH+I model, such as (i) obtaining a better fit in terms of log-likelihood value; (ii) checking for substitutional biases along individual edges of a phylogenetic tree; and (iii) estimating the nucleotide composition of invariant sites even when the process is nonstationary, nonreversible, and nonhomogeneous. The number of parameters for the BH+I model is much greater than that for the GTR+I model, and it may be argued that there is a risk of overfitting the data using models that are not based on biological knowledge. Using the parametric bootstrap, we showed that the GTR+I model is not sufficient to fully explain the complexity of the bacterial data set. The pivotal question, therefore, is how many parameters are needed to model the evolution of a set of sequences.

It should also be borne in mind that a good fit between the model and the data only indicates that the model is consistent with the data. In the present case, a good fit was obtained between the data and a model that does not consider the ribosomal RNA molecule's secondary structure (which implies that many of the sites most likely are not independent). The BH+I model can also be used as a guide for identifying simpler models that appropriately explain the variation in sequence data. For example, our analysis of the $Q$-matrices showed that the bacterial data set could be described using two different Markovian processes and such a Markov model (with fewer parameters than the BH+I model) was proposed by Foster (2004). However, another data set might require a more parameter-rich model such as the BH+I model.

Another direction of future research would be to extend the BH+I model to include information on rate heterogeneity across variable sites. Although the EM algorithm is computationally slower than that described above, it could be extended to incorporate rate heterogeneity across sites; however, it would have an extremely large number of parameters and might be impractical from a computational point of view.

A final area of future research would be to partition the computational problem associated with these parameter-rich models such that programs implementing them can be executed on supercomputers or...
distributed systems of processors. Recent papers have shown that it is possible to obtain a near-linear speedup of phylogenetic analyses on such computational resources (Keane et al., 2005; Zhou et al., 2007).

ACKNOWLEDGMENTS

We thank Mike Steel, Stephane Guindon, Derek Zwickl, and an anonymous reviewer for their constructive comments and suggestions. This research was partly funded by Discovery Grants (DP0453173 and DP0556820) from the Australian Research Council.

REFERENCES


First submitted 2 May 2006; reviews returned 9 July 2006; final acceptance 12 September 2006

Associate Editor: Mike Steel