Mixture Models of Nucleotide Sequence Evolution that Account for Heterogeneity in the Substitution Process Across Sites and Across Lineages

Vivek Jayaswal1,2, Thomas K.F. Wong3, John Robinson2,4, Leon Poladian2,4, and Lars S. Jermiin3,*

1School of Biomedical Sciences, Queensland University of Technology, Brisbane, QLD 4000, Australia; 2School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia; 3CSIRO Ecosystem Sciences, Canberra, ACT 2601, Australia; and 4Centre for Mathematical Biology, University of Sydney, Sydney, NSW 2006, Australia

*Correspondence to: CSIRO Ecosystem Sciences, Canberra, ACT, 2601, Australia; E-mail: lars.jermiin@csiro.au.

Received 25 March 2013; reviews returned 31 May 2013; accepted 10 May 2014
Associate Editor: Edward Susko

Abstract.—Molecular phylogenetic studies of homologous sequences of nucleotides often assume that the underlying evolutionary process was globally stationary, reversible, and homogeneous (SRH), and that a model of evolution with one or more site-specific and time-reversible rate matrices (e.g., the GTR rate matrix) is enough to accurately model the evolution of data over the whole tree. However, an increasing body of data suggests that evolution under these conditions is an exception, rather than the norm. To address this issue, several non-SRH models of molecular evolution have been proposed, but they either ignore heterogeneity in the substitution process across sites (HAS) or assume it can be modeled accurately using the Γ distribution. As an alternative to these models of evolution, we introduce a family of mixture models that approximate HAS without the assumption of an underlying predefined statistical distribution. This family of mixture models is combined with non-SRH models of evolution that account for heterogeneity in the substitution process across lineages (HAL). We also present two algorithms for searching model space and identifying an optimal model of evolution that is unlikely to over- or underparameterize the data. The performance of the two new algorithms was evaluated using alignments of nucleotides with 10 000 sites simulated under complex non-SRH conditions on a 25-tipped tree. The algorithms were found to be very successful, identifying the correct HAL model with a 75% success rate (the average success rate for assigning rate matrices to the tree’s 48 edges was 99.25%) and, for the correct HAL model, identifying the correct HAS model with a 98% success rate. Finally, parameter estimates obtained under the correct HAL-HAS model were found to be accurate and precise. The merits of our new algorithms were illustrated with an analysis of 42 337 second codon sites from a concatenation of 16 alignments of orthologous genes encoded by the nuclear genomes of Saccharomyces cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, S. castelli, S. kluyveri, S. bayanus, and Candida albicans. Our results show that second codon sites in the ancestral genome of these species contained 49.1% invariable sites, 39.6% variable sites belonging to one rate category (Γ1), and 11.3% variable sites belonging to a second rate category (Γ2). The ancestral nucleotide content was found to differ markedly across these three sets of sites, and the evolutionary processes operating at the variable sites were found to be non-SRH and best modeled by a combination of eight edge-specific rate matrices (four for Γ1 and four for Γ2). The number of substitutions per site at the variable sites also differed markedly, with sites belonging to Γ1 evolving slower than those belonging to Γ2 along the lineages separating the seven species of Saccharomyces. Finally, sites belonging to Γ2 appeared to have ceased evolving along the lineages separating S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, and S. bayanus, implying that they might have become so selectively constrained that they could be considered invariable sites in these species. [Evolution; heterotachy; mixture model; non-homogeneous model; phylogeny; rate heterogeneity across sites; rate heterogeneity across lineages; yeast]

Phylogenetic analyses are usually done using alignments of nucleotides or amino acids obtained from single or multiple genes. Compared to single-gene data, multigene data offer the advantage of being associated with lower stochastic errors (Phillips et al. 2004; Delsuc et al. 2005; Holland et al. 2006), which, in turn, entails a more consistent phylogenetic signal (see e.g., Jermiin et al. 2005) and an improved ability to discriminate between competing tree topologies (e.g., Strimmer and Rambaut 2002; Shi et al. 2005). On the other hand, systematic error caused by, for example, compositional heterogeneity across sequences and/or sites in alignments of nucleotides or amino acids (e.g., Gowri-Shankar and Rattray 2006; Ho et al. 2006) and long branch attraction (Felsenstein 1978) may be more pronounced in multigene data sets (e.g., Phillips et al. 2004; Delsuc et al. 2005). Compositional heterogeneity across sequences appears to be fairly widespread (for a compilation of examples, see Jermiin et al. 2009), implying that the data cannot have evolved under globally stationary, reversible, and homogeneous (SRH) conditions (for definitions, see Bryant et al. 2005; Jayaswal et al. 2005; Alabern et al. 2006a, 2006b; Jermiin et al. 2008). Computer simulations have furthermore shown that assuming evolution under globally SRH conditions when this is not justified can lead to bias in phylogenetic estimates (e.g., Ho and Jermiin 2004; Jermiin et al. 2004). Hence, it is unwise to assume that molecular evolution of sequences can be modeled accurately using site-specific, time-reversible Markov models (e.g., the GTR model by Lanave et al. 1984) when there is compositional heterogeneity across the sequences.

A non-homogeneous model of evolution explicitly accounts for heterogeneity in the substitution process across lineages (HAL) and several such models have been proposed in the literature (e.g., Yang and Roberts 1995; Galtier and Gouy 1998; Galtier et al. 1999; Foster 2004; Jayaswal et al. 2005; Blanquart and Lartillot 2006; Jayaswal et al. 2007; Blanquart and Lartillot 2008; Dutheil and Boussau 2008; Jayaswal et al. 2011b; Dutheil et al. 2012; Zou et al. 2012; Groussin et al. 2013). Several of these
models (e.g., Jayaswal et al. 2005, 2007, 2011b; Zou et al. 2012) assume a distinct rate matrix on each edge of the phylogenetic tree and are likely to overparameterize the data. Similarly, the N2 model by Yang and Roberts (1995) assumes a distinct vector of equilibrium nucleotide frequencies on each edge of the tree and is likely to overparameterize the data. Groussin et al. (2013) modified the N2 model for analysis of alignments of amino acids, but instead of considering a distinct vector of equilibrium amino acid frequencies on each edge, they considered linear combinations of these frequencies that best explain the compositional heterogeneity across the sequences. Although their approach reduces the number of free parameters, the underlying assumption is still that each edge is associated with a distinct rate matrix. Foster (2004) showed that model complexity could be reduced by allowing the same rate matrix to be assigned to two or more edges. A limitation of this method is that the number of distinct rate matrices has to be specified a priori and is not estimated from the data. By contrast, the Bayesian method by Blanquart and Lartillot (2006) does not have this limitation—instead, it employs the Poisson distribution to determine the number of changes (breakpoints) over the tree and, thus, the number of distinct rate matrices over the tree. Their method uses a set of carefully tuned proposal functions (for a theoretical description of such functions, see Chib and Greenberg 1995) to generate a Markov chain whose stationary distribution is the posterior probability distribution of the model’s parameters. Because the Blanquart and Lartillot (2006) method relies on a Markov Chain Monte Carlo algorithm, it is computationally demanding and was used by the authors to compare a relatively small number of taxa.

Recently, two non-Bayesian algorithms were proposed for estimating the optimal and near-optimal HAL models, given a tree (Jayaswal et al. 2011a; Dutheil et al. 2012). The algorithm by Dutheil et al. (2012) starts with a simple model of evolution (i.e., a model with the same rate matrix on all edges) and gradually increases the complexity of the model. Initially, the simplest model (i.e., one with the same rate matrix assigned to each edge of the phylogenetic tree) and then gradually increases the complexity of the model by assigning different rate matrices to different edges. In this article, we propose a similar algorithm, and illustrate its usefulness using simulated and real data sets. Our algorithm differs from that proposed by Dutheil et al. (2012) by not relying on substitution mapping. A rooted B-taxon tree has $2B - 2$ edges, implying that the number of distinct rate matrices can vary between 1 and $2B - 2$. If an instantaneous rate matrix, $R$, is specified as in Lanave et al. (1984), then it can be represented as $R_{ij} = S_{ij} \pi_j$, for $i \neq j$ and $R_{ii} = -\sum_j R_{ij}$, where $S_{ij} = S_{ji}$ and $i, j \in \{A, C, G, T\}$. Henceforth, $\pi$ denotes the vector of equilibrium frequencies and $S = \{S_{ij}\}$. Now, under the assumption that all sites are independent and identically distributed, the number of free parameters varies between $2B - 9$ and $18B - 15$. To model the evolution of, for example, multi-gene data sets, one may also wish to consider heterogeneity in the substitution process across sites (HAS). To do so, sites in the alignment need to be grouped into K sets and a HAL model is then applied to each set of variable sites. The number of free parameters for a HAL-HAS model now varies between $(2B + 10)K - 1$ and $(18B - 14)K - 1$. One way to reduce the number of free parameters is to assume that the K sets of sites have common $S$, $\pi$, and $f_0$ (i.e., the vector of nucleotide frequencies at the root), and that the edge lengths are scalar multiples of one another. Specifically, if the scalar multiples are obtained using the discretized $\Gamma$ distribution (Yang 1994), then the number of free parameters varies between $2B + 10$ and $18B - 14$. Although the $\Gamma$ distribution is the most popular choice for modeling these scalar multiples (e.g., Galtier and Gouy 1998; Galtier et al. 1999; Foster 2004; Blanquart and Lartillot 2008), other distributions (e.g.,
normal, beta, and exponential) can also be used (e.g., Gueguen et al. 2013). However, no biological reason justifies modeling HAS using scalar multiples of edge lengths, so biologically realistic models of HAS are clearly needed. One such model was proposed by Pagel and Meade (2004) under the assumption that there is no HAL. Because there is evidence of HAL in many single-gene data sets (e.g., Jayaswal et al. 2005, 2007, 2011a, 2011b), there is a need for models that consider both HAL and HAS. Moreover, the HAS models should consider sources of site-to-site variation other than those based on edge lengths. In this context, Blanquart and Lartillot (2008) extended their earlier model of HAL (Blanquart and Lartillot 2006) by allowing different categories of sites to have different equilibrium frequencies. Their new model of evolution assumes that the number of rate categories, $K$, is fixed and that all the sites, irrespective of the rate category to which they belong, have the same HAL model. Although their rate matrix has the same form as that proposed by Lanave et al. (1984), their model only considers $\pi$ when modeling the changes in rate matrices.

Here we introduce a family of HAL-HAS models that considers, on the one hand, more general conditions than those assumed by Blanquart and Lartillot (2008), and, on the other, a simple mixture model equivalent to the $\Gamma$-distribution-based model of HAS. In the following section, we describe two new algorithms for identifying the optimal HAL model and, for a given HAL model, for identifying the optimal HAS model. We first apply our new algorithms to a 25-taxa data set simulated under complex non-SRH conditions and show that they identify the correct model of evolution in the vast majority of cases. Next, we apply our algorithms to a phylogenomic data set generated by concatenating alignments of 106 nuclear, protein-coding genes from eight species of yeast. These data were first analyzed by Rokas et al. (2003) and were subsequently found to be compositionally heterogeneous (Phillips et al. 2004; Collins et al. 2005), although the nature of this heterogeneity has not yet been fully elucidated. Using our algorithms, we found a HAL-HAS model that fits the data better than the other models considered. In addition, the parameter values for this model disclosed a complex non-SRH evolutionary process.

### Methods

Let $X$ be a $B \times N$ matrix of $B$ taxa and $N$ homologous nucleotide sites, and let $T$ be a rooted, binary tree with $2B-2$ edges. We assume that the Markov process over an edge in $T$ is time-homogeneous and can be represented by a $4 \times 4$ instantaneous rate matrix, $R = \{R_{ij}\}$, as described in the Introduction Section, where $i, j = 1, 2, 3,$ and 4 represent $A, C, G,$ and $T,$ respectively. By equating two or more of the elements in $R$ (i.e., $S_{ij}$ and $\pi_{i}$), we obtain simpler rate matrices, such as the HKY model (Hasegawa et al. 1985). As it is impossible to estimate both $R$ and the edge length (denoted by $t$) simultaneously, we impose the constraint $-\sum_{i=1}^{4} \pi_{i} R_{ii} = 1$ on the rate matrix and estimate $t$.

#### Modeling HAL

For a given tree, $T$, a HAL model, denoted $W$, consists of a partition $G$ of edges, $E$, such that the number of elements in $G$ (denoted $G = |G|$) varies between 1 and $2B-2$. The model parameters, $\theta_{W}$, include:

1. the edge lengths $l_{b}$, where $b \in E$ and $t$ denotes the expected number of substitutions per site;
2. the vector of frequencies at the root (i.e., $f_{0}$); and
3. the distinct rate matrices $R_{g}$, $g \in G$. We note that all edges in the subset $g$ are assigned the rate matrix $R_{g}$.

Let $T_{l}$ and $T_{r}$ denote, respectively, the left and right subtrees obtained by splitting $T$ at the root, and let $E_{l}$ and $E_{r}$, respectively, denote the edges of $T_{l}$ and $T_{r}$. For the $n$-th site, we denote the nucleotides at the leaf nodes of subtrees $T_{l}$ and $T_{r}$ by $X_{n}^{l}$ and $X_{n}^{r}$, respectively. Now, for a given HAL model, $W$, the likelihood of the $n$-th site is

$$ L_{W} = \frac{4}{\sum_{i=1}^{4} f_{0,i} P(X_{n}^{l}|\theta_{W}, i)P(X_{n}^{r}|\theta_{W}, i).} $$

where $f_{0}$ is the $i$-th element of $f_{0} = [f_{01}, f_{02}, f_{03}, f_{04}]^{T}$, and $P(X_{n}^{l}|\theta_{W}, i)$ and $P(X_{n}^{r}|\theta_{W}, i)$ denote the conditional probabilities of obtaining the nucleotides in $X_{n}^{l}$ and $X_{n}^{r}$, respectively, under the model with parameters $\theta_{W}$.

#### Modeling HAS

HAS is modeled under the assumption that the sites in an alignment can be grouped into $K > 1$ categories and that all the site categories have the same HAL model, $W$. Now, for a given category $k \in K$, $\theta_{k}$ denotes the set of model parameters (including $p_{k}^{l}$, which denotes the vector of nucleotide frequencies at the root for the $k$-th category) and the likelihood of the $n$-th site is

$$ L_{k}^{W} = \frac{4}{\prod_{i=1}^{K} \sum_{k=1}^{K} \sum_{i=1}^{4} f_{k,i} \prod_{j=1}^{4} P(X_{n}^{j}|\theta_{k}, j)P(X_{n}^{j}|\theta_{W}, j),} $$

where $q_{k}$ denotes the probability that a site belongs to the $k$-th rate category and $\sum_{k=1}^{K} q_{k} = 1$. Under the assumption of independence of the sites, the overall likelihood is

$$ L_{W} = \prod_{n=1}^{N} L_{n}^{W} $$
to as invariable sites. To accommodate the potential prevention of some sites from evolving. These sites are referred the three-dimensional structures of proteins and RNAs by real data because the topological constraints on grouped into

<table>
<thead>
<tr>
<th>Model</th>
<th>Common $f_3$</th>
<th>Common $S$</th>
<th>Common $\pi$</th>
<th>Scalar edge lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAS1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>HAS2</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>HAS3</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>HAS4</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>HAS5</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The five HAS models can be defined in a similar manner for $K > 2$.

Identification of the optimal HAL model

Since every unique partition of edges represents a HAL model, the total number of HAL models is a Bell number (Jayaswal et al. 2011a). For example, the number of models is $190,899,322$ for a rooted, eight-taxon tree, so an exhaustive search of all models is infeasible. Subsequently, we present two algorithms to identify the optimal and/or near-optimal HAL models.

Algorithm 1: model generation

A HAL model consists of a non-overlapping clustering of edges into $G$ groups such that each group is assigned a distinct rate matrix $R_{g} \in G$ (see subsection “Modelling HAL”). In this subsection, we treat $R_1, \ldots, R_G$ as labels for clustering the edges and note that, for a given clustering, the estimates of a model’s parameters are obtained using the maximum-likelihood approach.

Input parameters:

1. A model of evolution $W$
2. An internal node $d$

Output: Four new models, which are more complex than the original model, $W$

Steps:

1. Obtain $\mathcal{R}$ (i.e., the set of all unique rate matrices in $W$). Generate two new rate matrices $R_1$ and $R_2$ such that $R_1 \neq R_2$ and $R_1, R_2 \notin \mathcal{R}$.
2. For the sub-tree rooted at node $d$ (i.e., $T_d$), perform the following steps—
   (i) Identify the internal nodes for which at least one of the child edges does not have the same rate matrix as the parent edge.

Invariable sites

So far, we have assumed that the $N$ sites can be grouped into $K$ categories and that all the sites within a category are variable. This assumption is often violated by real data because the topological constraints on the three-dimensional structures of proteins and RNAs prevent some sites from evolving. These sites are referred to as invariable sites. To accommodate the potential presence of invariable sites in the context of our HAL-HAS model, we modified Equation 2 to incorporate these sites as follows:

$$L_W = \sum_{i=1}^{4} \sum_{k=1}^{K} f_k^i \pi^{i_0} + \sum_{k=1}^{K} f_k^i \sum_{l=1}^{4} P(X^{i_0} \mid \theta_k^i, \pi) P(X^{i_0} \mid \theta_k^i, \pi)$$

(5)

where $\pi$ denotes the probability that a site is invariable, $\beta + \sum_{k=1}^{K} a_k = 1$, and $f_k^i$ is an indicator function that takes the value 1 if all the taxa have nucleotide $i$ at the $h$-th site and 0, otherwise. Also, $\pi^{i_0}$, where $i = 1, 2, 3, 4$, is the conditional probability of observing nucleotide $i$ in all the taxa, given that the site is invariable. Unlike the commonly used approach of modeling invariable sites, we do not assume that $\pi^{i_0}$ is equal to the vector of equilibrium nucleotide frequencies for the variable sites (Jayaswal et al. 2011a; 2011b).

Identification of the optimal HAL model

Since every unique partition of edges represents a HAL model, the total number of HAL models is a Bell number (Jayaswal et al. 2011a). For example, the number of models is $190,899,322$ for a rooted, eight-taxon tree, so an exhaustive search of all models is infeasible. Subsequently, we present two algorithms to identify the optimal and/or near-optimal HAL models.

Algorithm 1: model generation

A HAL model consists of a non-overlapping clustering of edges into $G$ groups such that each group is assigned a distinct rate matrix $R_g \in G$ (see subsection “Modelling HAL”). In this subsection, we treat $R_1, \ldots, R_G$ as labels for clustering the edges and note that, for a given clustering, the estimates of a model’s parameters are obtained using the maximum-likelihood approach.

Input parameters:

1. A model of evolution $W$
2. An internal node $d$

Output: Four new models, which are more complex than the original model, $W$

Steps:

1. Obtain $\mathcal{R}$ (i.e., the set of all unique rate matrices in $W$). Generate two new rate matrices $R_1$ and $R_2$ such that $R_1 \neq R_2$ and $R_1, R_2 \notin \mathcal{R}$.
2. For the sub-tree rooted at node $d$ (i.e., $T_d$), perform the following steps—
   (i) Identify the internal nodes for which at least one of the child edges does not have the same rate matrix as the parent edge.
Algorithm 2: search for HAL model

Assumptions: The Search for HAL model (SHAL) algorithm assumes that the sites are either variable or invariable, and that all the variable sites belong to the same rate category (i.e., $K = 1$).

Input parameters: Tree topology $T$

Output: A set of HAL models

Steps:

I. Initialization step: we denote a set of HAL models $\mathcal{S}$, a user-defined threshold value by $J$, and the current iteration of the algorithm by $a$. We denote the simplest HAL model, with the same rate matrix on each edge, by $W_1$. In general, we denote by $W_a$ the model being evaluated during the $a$-th iteration. We initialize the SHAL algorithm by setting $S = \{W_1\}$, $J = 3$, and $a = 1$.

II. Increase model complexity (bottom-up approach): in this step of the algorithm, we use the current model $W_a$ to identify more complex models that fit the data better than $W_a$ using an information-theoretic criterion (ITC). The ITC could be either AIC or BIC. The substeps corresponding to this step of the algorithm are:

(i) Obtain the set $U$ of internal nodes such that the parent edge and the two child edges have a common rate matrix. For the root node, there is no parent edge, so include it in $U$ if the two child edges have the same rate matrix.

(ii) Identify all the descendant edges (not just the child edges) corresponding to these internal nodes. Refer to this set of edges as $\mathcal{E}_U$ and the remaining edges (in $T_d$) as $\mathcal{E}_d$ (Fig. 1a).

(iii) Generate a new model $W_1$ from $W$ by assigning $R_1$ to all the edges in $\mathcal{E}_d$ (Fig. 1b).

(iv) Generate a new model $W_2$ from $W$ by assigning $R_1$ to edges that belong to $\mathcal{E}_U$ in the left subtree of $T_d$ (Fig. 1c).

(v) Generate a new model $W_3$ from $W$ by assigning $R_1$ to edges that belong to $\mathcal{E}_U$ in the right subtree of $T_d$ (Fig. 1d).

(vi) Generate a new model $W_4$ from $W$ by assigning $R_1$ and $R_2$ to the edges that belong to $\mathcal{E}_U$ in the left and right subtrees of $T_d$, respectively (Fig. 1e).

III. Return the new models

The revised models $W_1, \ldots, W_4$ have more parameters than $W$ because some of the edges in tree $T$ have the new rate matrices associated with them.

Algorithm 2: search for HAL model

Assumptions: The Search for HAL model (SHAL) algorithm assumes that the sites are either variable or invariable, and that all the variable sites belong to the same rate category (i.e., $K = 1$).

Input parameters: Tree topology $T$

Output: A set of HAL models

Steps:

I. Initialization step: we denote a set of HAL models $\mathcal{S}$, a user-defined threshold value by $J$, and the current iteration of the algorithm by $a$. We denote the simplest HAL model, with the same rate matrix on each edge, by $W_1$. In general, we denote by $W_a$ the model being evaluated during the $a$-th iteration. We initialize the SHAL algorithm by setting $S = \{W_1\}$, $J = 3$, and $a = 1$.

II. Increase model complexity (bottom-up approach): in this step of the algorithm, we use the current model $W_a$ to identify more complex models that fit the data better than $W_a$ using an information-theoretic criterion (ITC). The ITC could be either AIC or BIC. The substeps corresponding to this step of the algorithm are:

(i) Obtain the set $U$ of internal nodes such that the parent edge and the two child edges have a common rate matrix. For the root node, there is no parent edge, so include it in $U$ if the two child edges have the same rate matrix.
ii) For every internal node, \( d \in \mathcal{U} \), generate new models using Algorithm 1 with input parameters \( W_d \) and \( d \).

iii) Sort all models (i.e., \( 4 \times |\mathcal{U}| \) models) generated in Step (ii) in ascending order of ITC values and identify the top \( J \) models. Label these models as \( W_d^j \), \( j = 1, \ldots, J \). Add a model \( W_d^J \) to the set \( S \) if the ITC value for this model is lower than that for the original model \( W_d \).

iv) Set \( a = a + 1 \). If the number of elements in \( S \) is \( \geq a \), then repeat Steps (i–iv) using the revised \( a \)-th model \( W_d^a \). Otherwise go to Step (v).

v) Sort the models in \( S \) in increasing order of ITC values. The first model (i.e., the model with the lowest ITC value) is denoted as \( W_d^{aW} \). This model represents the best HAL model identified using a stepwise increase in model complexity (here, BU refers to “bottom-up”).

III. Decrease model complexity (top-down approach): to determine whether \( W_d^{aW} \) is an overparameterized model of evolution, we apply the CORE algorithm of Jayaswal et al. (2011b) to \( W_d^{aW} \). During each step of the CORE algorithm, we first identify the number of unique rate matrices, say \( G \). Second, we cluster the \( \pi \) vectors associated with the \( G \) rate matrices and obtain the two closest vectors. Last, we equate the rate matrices corresponding to the two closest \( \pi \) vectors, thereby, obtaining a simpler model. This stepwise reduction in model complexity continues until all the edges are assigned the same rate matrix (i.e., until \( G = 1 \)). Add the set of models evaluated in this step to \( S \).

IV. Termination step: once the CORE algorithm has finished execution, return all the models in set \( S \).

We sort the output HAL models in increasing order of ITC values and obtain the optimal (denoted \( W_{\text{opt}} \)) and near-optimal models.

Identification of the optimal HAS model

Once the optimal HAL model has been identified, we consider models with multiple categories (i.e., \( K > 1 \)) of variable sites (described in subsection “Modeling HAS”). We increase \( K \) from 2 to a user-defined threshold value, say \( K^{\text{max}} \), and for each value of \( K \) evaluate all the five HAS models (Table 1). Table 2 shows the number of free parameters associated with each of these HAL-HAS models. We use the ITC value to identify the optimal combination of \( K \) and HAS model (i.e., the one that best fits the data). We refer to this model as the optimal HAL-HAS model as it takes into account rate heterogeneity across sites and across lineages.

A comment on the use of the AIC or BIC for model selection

Model selection that relies on the AIC or BIC is subject to biases in the same way that model-selection procedures in stepwise regression are, for the same reasons as outlined in Section 5 of Kadane and Lazar (2004). However, the suitability of the AIC or BIC for finding a near-optimal HAL-HAS model is supported by simulation studies described later in the article.

Software

The algorithms described above were implemented as C++ programs (HAL-BU, HAL-TD, and HAS). Here, HAL-BU, HAL-TD, and HAS refer, respectively, to the bottom-up approach, the top-down approach (i.e., the CORE algorithm; formerly written in R), and the algorithm used to detect heterogeneity in substitutions across sites, given a HAL model. The three programs are available from http://www.csiro.au/hal-has. Sequences generated in silico as part of this study were generated using Hetero version 2.0, which is available from http://www.csiro.au/hetero2.

RESULTS AND DISCUSSION

We first evaluated the accuracy of our algorithms in identifying the true HAL-HAS model using simulated data. Second, we applied our algorithms to an eight-taxon yeast data set and showed that the HAL-HAS model identified using our algorithms fits the data better than models obtained using other currently available methods. The initial parameter values for our algorithms are described in Appendix 1(a) and 1(b). In the next subsections of the article, values after ‘\( \pm \)’ represent sample standard deviations (SDs).

<table>
<thead>
<tr>
<th>Model</th>
<th>Edges</th>
<th>( S )</th>
<th>( \pi )</th>
<th>( \pi^{\text{ref}} )</th>
<th>( f_0 )</th>
<th>( x^a )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAL-HAS1</td>
<td>(2F-2K)</td>
<td>5GK</td>
<td>3GK</td>
<td>3</td>
<td>3K</td>
<td>K</td>
<td>2BK + 8GK + 2K + 3</td>
</tr>
<tr>
<td>HAL-HAS2</td>
<td>(2F-2K)</td>
<td>5G</td>
<td>3G</td>
<td>3</td>
<td>3K</td>
<td>K</td>
<td>2BK + 3G + 5G + 2K + 3</td>
</tr>
<tr>
<td>HAL-HAS3</td>
<td>(2F-2K)</td>
<td>5G</td>
<td>3G</td>
<td>3</td>
<td>3K</td>
<td>K</td>
<td>2BK + 8G + 2K + 3</td>
</tr>
<tr>
<td>HAL-HAS4</td>
<td>(2F-2K)</td>
<td>5G</td>
<td>3G</td>
<td>3</td>
<td>3K</td>
<td>K</td>
<td>2BK + 8G - K + 6</td>
</tr>
<tr>
<td>HAL-HAS5</td>
<td>(2F-2K-1)</td>
<td>5G</td>
<td>3G</td>
<td>3</td>
<td>3K</td>
<td>K</td>
<td>2BK + 8G + 2K + 3</td>
</tr>
</tbody>
</table>

*Probability that a site belongs to a given category.
Table 3. Actual and inferred proportions of invariant sites (π), variable sites belonging to V1 (a1), and variable sites belonging to V2 (a2) for simulated data.

<table>
<thead>
<tr>
<th>Type</th>
<th>β</th>
<th>a1</th>
<th>a2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>0.49674</td>
<td>0.35473</td>
<td>0.14853</td>
</tr>
<tr>
<td>Inferred</td>
<td>0.49669 ± 0.00007</td>
<td>0.35621 ± 0.00078</td>
<td>0.14710 ± 0.00877</td>
</tr>
</tbody>
</table>

*Values after ± indicate the sample SD.

Generation of simulated data

We generated 100 data sets, each comprising 25 nucleotide sequences and 10,000 sites, on a rooted 25-taxon tree. We considered three categories of sites (variable sites and two categories of variable sites, viz., V1 and V2) and, for variable sites, four edge-specific rate matrices. Each data set was obtained using the following procedure:

(a) An ancestral sequence was generated using the parameter values specified in Tables 3 and 4 for variable sites (β, πinv), V1 (a1, f1), and V2 (a2, f2);
(b) The ancestral sequence was allowed to evolve on the taxon tree. We considered three categories of sites (V1, V2, V3) and, for variable sites, four edge-specific rate matrices. Each data set was obtained using the following procedure:

c) The edge lengths were set to differ for V1 and V2, as shown in Figures 2a and 2b.

Overall, the evolutionary process for the simulated data corresponded to a HAL-HAS3 model and the parameter values were chosen so that they were similar to those obtained from the yeast data for the HAL-HAS3 model.

Assessment of the HAL algorithm’s performance

We analyzed the simulated data using the SHAL algorithm to assess its performance. For a 25-taxon tree, the total number of HAL models (assuming the same HAL model for all the sites) is 6.3 x 10^44. On average, our algorithm compared 2400 ± 218 models before returning the optimal model. The optimal model was correct in 75% of the cases and always included four rate matrices. These results demonstrate that when the data are generated under the conditions allowed for by our HAL-HAS models, our algorithm is able to identify the true HAL model by surveying a tiny proportion of the total set of models available.

We also surveyed those cases where the optimal model was incorrect to better understand the nature of the model misspecification. In those cases, an incorrect rate matrix was assigned to 1 (16 cases), 2 (7 cases), or 3 (2 cases) of the 48 edges, and only 4 of the 48 edges were associated with a misspecified rate matrix (i.e., e31, e32, e45, and e48; Fig. 2a). In other words, when considering the assignment of rate matrices to the individual edges (irrespective of whether the HAL model is correct or not), the rate of success was 99.25%.

The above-mentioned four edges are found close to the root of the tree and are, in two cases, quite short (i.e., e31 and e32 in Fig. 2b). However, the assignment of rate matrices to other short edges (e.g., e29 and e32; Fig. 2b; e29, and e32; Fig. 2c) was always correct, perhaps because some of these edges are found closer toward the tips of the tree (e.g., e29, and e32; Fig. 2c), indicating, as expected, that assigning a correct rate matrix to a short edge close to the root of the tree may be more difficult than doing so for a short edge close to the tips of the tree.

Assessment of the HAS algorithm’s performance

We assumed the true HAL model for each of the 100 simulated data sets and analyzed the data using HAS. Given that the simulated data were obtained under a HAS3 model with two categories of variable sites (i.e., K = 2), we examined whether our algorithm is accurate, or tends to over- or underparameterize the data.

Table 4. Actual and inferred nucleotide frequencies for the invariant sites (πinv), the root vectors (f0), and the rate matrices (f) for simulated data.

<table>
<thead>
<tr>
<th>Type</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>0.30551</td>
<td>0.16659</td>
<td>0.13522</td>
<td>0.39268</td>
</tr>
<tr>
<td>Inferral</td>
<td>0.30564 ± 0.00009</td>
<td>0.16650 ± 0.00003</td>
<td>0.13529 ± 0.00002</td>
<td>0.39260 ± 0.00006</td>
</tr>
<tr>
<td>f1</td>
<td>0.23175</td>
<td>0.21523</td>
<td>0.27974</td>
<td>0.27980</td>
</tr>
<tr>
<td>f2</td>
<td>0.22893 ± 0.00051</td>
<td>0.21462 ± 0.00139</td>
<td>0.28267 ± 0.00201</td>
<td>0.27488 ± 0.00105</td>
</tr>
<tr>
<td>f3</td>
<td>0.48586</td>
<td>0.15915</td>
<td>0.15888</td>
<td>0.19831</td>
</tr>
<tr>
<td>f4</td>
<td>0.49863 ± 0.02866</td>
<td>0.15540 ± 0.00858</td>
<td>0.16463 ± 0.02887</td>
<td>0.19934 ± 0.01360</td>
</tr>
<tr>
<td>x1</td>
<td>0.50704</td>
<td>0.23941</td>
<td>0.19537</td>
<td>0.21428</td>
</tr>
<tr>
<td>x2</td>
<td>0.30666 ± 0.00048</td>
<td>0.21781 ± 0.00067</td>
<td>0.19365 ± 0.00416</td>
<td>0.21248 ± 0.00659</td>
</tr>
<tr>
<td>x3</td>
<td>0.33362</td>
<td>0.26804</td>
<td>0.16882</td>
<td>0.22952</td>
</tr>
<tr>
<td>x4</td>
<td>0.33529 ± 0.00061</td>
<td>0.26599 ± 0.00055</td>
<td>0.16791 ± 0.00528</td>
<td>0.21241 ± 0.00728</td>
</tr>
</tbody>
</table>

*Values after ± indicate the sample SD.
The optimal HAS model was incorrect, the optimal model a 98% rate of success. In both the two cases where the true model for all but 2 of the 100 data sets, implying models. The optimal HAS model corresponded to the V and very similar. Table 5 shows that the true and inferred equilibrium frequencies for the rate matrices were also edge lengths for the two categories of variable sites (V1). (c) The tree, with its edges drawn to scale, for the other set of variable sites (V2). The scale bar corresponds to 0.1 substitution per site.

We considered K = 2 and 3 and evaluated all five HAS models. The optimal HAS model corresponded to the true model for all but 2 of the 100 data sets, implying a 98% rate of success. In both the two cases where the optimal HAS model was incorrect, the optimal model was found to be HAS4 with K = 2, implying a slight tendency to underparameterize the data.

Accuracy of the HAL-HAS model’s parameter estimates

To evaluate the accuracy of parameter estimates for the true HAL-HAS model, we calculated the average and sample SD for each parameter’s estimate over the 100 simulated data sets. Table 3 shows that the true and inferred proportions of sites in various categories were very close. Table 4 shows that the true and inferred equilibrium frequencies for the rate matrices were also very similar. Table 5 shows that the true and inferred values of elements in S were very close, but the SD of the elements in S were larger than those for other variables. Overall, these results imply that the parameter estimates are accurate when the true model is used. Figure 3 displays two scatter plots with the actual and inferred edge lengths for the two categories of variable sites (V1 and V2). Apart from two outliers in each case, there is a good agreement between the inferred and actual edge lengths. These results imply that the estimates of edge lengths for different categories of variable sites are fairly accurate under the true HAL-HAS model.

Comparison of HAS models that consider scalar multiples of edge lengths

We now focus on HAS and assume that, within a rate category, all the edges have the same rate matrix. We compared the widely used GTR+Y model with a special case of the HAL-HAS model (i.e., an SRH HAL model with edge lengths that vary across rate categories and are scalar multiples of one another). We refer to the latter model as the GTR+Y model—Table 6 lists the key differences between these two models. If the number of rate categories (i.e., K) needs to be specified, then we use the notation GTR+\(X_K\) and GTR+\(Y_K\).

GTR+\(Y_2\) is the true model: we generated 100 data sets (each comprising 25 sequences with 10,000 nucleotides) under the GTR+\(Y_2\) model, using the parameters described in Appendix 1(c). Next, we analyzed each data set under the GTR+\(Y_2\) and GTR+\(Y_2\) models assuming the true tree topology. As expected, the BIC value for the GTR+\(Y_2\) model was always better than that for the
Table 5. Actual and inferred values of $S$ in simulated data

<table>
<thead>
<tr>
<th>Type</th>
<th>$S_{12}$</th>
<th>$S_{13}$</th>
<th>$S_{14}$</th>
<th>$S_{23}$</th>
<th>$S_{24}$</th>
<th>$S_{34}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^1$ Actual</td>
<td>0.58005</td>
<td>3.37592</td>
<td>0.53758</td>
<td>2.06688</td>
<td>0.24599</td>
<td></td>
</tr>
<tr>
<td>Inferred</td>
<td>0.53944 ± 0.04051</td>
<td>3.39962 ± 0.09933</td>
<td>0.52526 ± 0.02893</td>
<td>1.42401 ± 0.07124</td>
<td>2.04259 ± 0.07472</td>
<td>0.25340 ± 0.03962</td>
</tr>
<tr>
<td>$S^2$ Actual</td>
<td>1.32162</td>
<td>2.19708</td>
<td>0.84735</td>
<td>1.32205</td>
<td>0.42236</td>
<td></td>
</tr>
<tr>
<td>Inferred</td>
<td>1.31875 ± 0.04966</td>
<td>2.21266 ± 0.07763</td>
<td>0.84771 ± 0.03685</td>
<td>2.11223</td>
<td>0.42236</td>
<td></td>
</tr>
<tr>
<td>$S^3$ Actual</td>
<td>1.23980</td>
<td>1.63563</td>
<td>1.4410</td>
<td>1.62999</td>
<td>0.39482</td>
<td></td>
</tr>
<tr>
<td>Inferred</td>
<td>1.23872 ± 0.04530</td>
<td>1.68951 ± 0.05090</td>
<td>1.44663 ± 0.05581</td>
<td>1.63250 ± 0.05337</td>
<td>0.39339 ± 0.03656</td>
<td></td>
</tr>
<tr>
<td>$S^4$ Actual</td>
<td>1.4341</td>
<td>0.03990</td>
<td>1.97134 ± 0.17754</td>
<td>1.37466 ± 0.03421</td>
<td>0.05091</td>
<td>0.0509</td>
</tr>
<tr>
<td>Inferred</td>
<td>1.49552</td>
<td>2.97199</td>
<td>1.3775</td>
<td>9.32502</td>
<td>2.94626</td>
<td>3.64593</td>
</tr>
</tbody>
</table>

$^a$Values after ± indicate the sample SD.

Table 6. Differences between GTR+$\Gamma$ and GTR+$\Upsilon$ models

<table>
<thead>
<tr>
<th>Feature</th>
<th>GTR+$\Gamma$ model</th>
<th>GTR+$\Upsilon$ model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of sites in different rate categories ($\alpha_1, \cdots, \alpha_k$)</td>
<td>$\alpha_1 = \cdots = \alpha_k = 1/K$</td>
<td>$\alpha_1 + \cdots + \alpha_k = 1$</td>
</tr>
<tr>
<td>Scalar multiples of edge lengths ($\mu_1, \cdots, \mu_k$)</td>
<td>Obtained using a $\Gamma$ distribution</td>
<td>$\mu_1 = 1$ and the remaining scalar multiples are estimated</td>
</tr>
</tbody>
</table>

Identification of the actual rate category for a given site

We now focus on the use of a HAS model to identify regions, or groups, of contiguous sites that have evolved under the same condition, in a manner similar to that proposed by Pagel and Meade (2004). Initially, we generated a data set comprising 25 sequences with 10 000 nucleotides, under the GTR+$\Upsilon_2$ model, as stated in Appendix 1(d). In this particular case, the first 3 270 sites were assigned to $\nu_1$ while the remaining 6 730 sites were assigned to $\nu_2$. Next, we obtained $\Delta \log L = \log L_{\nu_1} - \log L_{\nu_2}$ for every site, where $\log L_{\nu_1}$ and $\log L_{\nu_2}$ denote the log-likelihoods under the parameter estimates corresponding to $\nu_1$ and $\nu_2$, respectively.

GTR+$\Upsilon_2$ is the true model: we generated 100 data sets (each comprising 25 sequences with 10 000 nucleotides) under the GTR+$\Upsilon_2$ model, using the parameters described in Appendix 1(d). Next, we analyzed each data set under the GTR+$\Gamma_2$ and GTR+$\Upsilon_2$ models for the true tree topology. As expected, the BIC value for the GTR+$\Upsilon_2$ model was always better than that for the GTR+$\Gamma_2$, and the difference was large, ranging from 1 000.2 to 1 108.8.

To check whether the GTR+$\Gamma$ model with higher values of $K$ might fit the data better, we also analyzed the data generated under the GTR+$\Upsilon_2$ model using GTR+$\Gamma_K$ models with $K = 3, \ldots, 12$. For each of the 100 data sets, irrespective of the $K$ value, the GTR+$\Gamma_K$ model always led to a poorer fit of the data than the GTR+$\Upsilon_2$ model. Moreover, within the GTR+$\Gamma_K$ family of models, the best model (in terms of its BIC value) was obtained for $K = 3$ in all but one case. Therefore, for data generated under the GTR+$\Upsilon_2$ model, the true model was always preferred over a GTR+$\Gamma_K$ model.
FIGURE 4. Distribution of $\Delta \log L = \log L_{V1} - \log L_{V2}$ along the alignment of the simulated data. If $\Delta \log L$ is positive, then the site is inferred as belonging to $V_1$; otherwise, the site is inferred as belonging to $V_2$. Figure 4 shows the distribution of $\Delta \log L$ values along the alignment. Of the 10,000 sites in this alignment, 3,145 sites (96.18%) were correctly inferred as belonging to $V_1$ while 6,630 sites (98.51%) were correctly inferred as belonging to $V_2$. Overall, 97.75% of the sites were assigned to the correct rate category, implying that our HAS model may be useful for assigning individual sites to specific rate categories. However, we note that the accuracy of this assignment will depend on the number of sequences, with more sequences likely to lead to more accurate estimates.

Analysis of a phylogenomic yeast data

We now consider a data set comprising 42,337 second codon position sites extracted from a concatenation of 106 alignments of orthologous genes encoded by the nuclear genes of S. cerevisiae (Scer), S. paradoxus (Spar), S. mikatae (Smik), S. kudriavzevi (Skud), S. castellii (Scas), S. kluyveri (Sklu), S. bayanus (Sbay), and C. albicans (Calb) (available as supplementary material from Dryad: http://dx.doi.org/10.5061/dryad.9hb68). These alignments were originally analyzed by Rokas et al. (2003) to determine how many genes are needed in a concatenation of aligned genes to resolve gene-specific incongruence in inferred molecular phylogenies. Subsequently, the data have been subject to extensive re-analysis (e.g., Phillips et al. 2004; Collins et al. 2005; Hedtke et al. 2006). Phillips et al. (2004) found evidence of compositional heterogeneity across the sequences and showed that it could mislead phylogenetic estimates. Collins et al. (2005) found that the data could be divided into genes that appear to have ($n = 54$), or have not ($n = 52$), evolved under stationary conditions; using these sets of genes, they then went on to discover that the stationary genes were superior to the non-stationary genes in terms of their ability to recover the assumed correct phylogeny. Hedtke et al. (2006) discovered that an additional source of conflict between the 106 genes might be branch-length heterogeneity. None of these studies examined the data using statistical tests designed for detecting compositional heterogeneity across sequences (reviewed in Jermiin et al. 2004). These tests include the matched-pairs test of symmetry (Bowker 1948), the matched-pairs test of marginal symmetry (Stuart 1955), and the matched-pairs test of internal symmetry (Ababneh et al. 2006a)—the formal relationship between these tests is described in Ababneh et al. (2006a) and Jermiin et al. (2008).

Preliminary survey of the yeast data

We performed the above-mentioned matched-pairs tests of homogeneity for all pairs of sequences in the data—for eight taxa, the total number equals 28. The PP plots shown in Figure 5 compare the observed and expected $P$-values obtained using the matched-pairs test of symmetry, marginal symmetry, and internal symmetry. In each case, the distribution of points, where each point represents a pair of sequences, is distinctly nonlinear, with 67.9% (symmetry), 64.3% (marginal symmetry), and 28.6% (internal symmetry) of the points $<0.05$. This implies that the data are highly unlikely to have evolved under globally SRH conditions (for details on how to interpret the results from these tests, see Jermiin et al. 2008; Vera-Ruiz et al. 2014). The PP plots also show that some of the pairs of sequences (i.e., those generating large observed $P$-values) may have evolved under similar conditions. This implies that different edge-specific rate matrices may be required to model the evolution of these sequences. The heat map shown in Figure 6 corroborates this observation, suggesting that a group comprising Scer, Spar, Smik, Skud, Sbay, and Scas may have evolved under similar conditions. The other two sequences appear to have evolved under mutually different conditions, implying that at least three rate
matrices may be needed to model the evolution of these sequences.

Optimal HAL model for the yeast data

To determine how best to model the evolution of these data, we decided to analyze them using the HAL-HAS family of models. In so doing, we used the tree topology shown in Figure 7. We chose this tree because it is one of the most widely accepted trees of yeast evolution (e.g., Kurtzman and Robnett 2003; Rokas et al. 2003). While other trees could also have been considered (e.g., those published in Rokas et al. 2003 and Phillips et al. 2004), the focus of the present article is on the identification of the optimal model of evolution for a given tree topology.

A rooted, eight-taxon tree has 14 edges, so for K = 1 the possible number of HAL models is 190,899,322. As it was not feasible to evaluate all of these models, we used the SHAL algorithm. Step II of this algorithm compared 666 models of evolution and identified an optimal model with four distinct rate matrices distributed over the 14 edges (i.e., \( W_{BU} \) with a BIC value of 263,713). Interestingly, Step III of this algorithm did not return a better model than \( W_{BU} \) so we concluded that \( W_{BU} \) is the optimal HAL model for these data.

We also applied the CORE algorithm to these data, starting from a model with 14 distinct rate matrices (i.e., one for each edge). The algorithm compared 133 models of evolution and found an optimal model with three rate matrices distributed over the 14 edges (i.e., \( W_{TD} \), with a BIC value of 263,789). As the BIC value for this model was much higher than that for \( W_{BU} \), we concluded that the latter model is the optimal HAL model for these data. The arrangement of rate matrices for this model, henceforth referred to as \( W_{opt} \), is shown in Figure 7.

Optimal HAS model for the yeast data

Next, we used \( W_{opt} \) to search for the optimal HAL-HAS model. We considered the HAS models shown in Table 1 and varied K between 2 and 4. For \( W_{opt} \), the best BIC value was obtained for the combination of HAS2 and K = 2 (Table 7). Henceforth, we refer to the combination of \( W_{opt} \), HAS2, and K = 2 as the optimal HAL-HAS model for these data. We note that, for the yeast data, each of the HAS models had the best BIC value for K = 2 (Table 7), which is why we did not consider values of K > 4. Finally, we note that the HAL-HAS models had multiple maxima, and that the log-likelihood and BIC values presented in this section of the study were obtained using the initial parameter values shown in Appendix 1(a) and 1(b).

Comparing optimal models of evolution for the yeast data

We compared \( H_{opt} \) to the GTR+I+Γ model, which assumes evolution under globally SRH conditions—an
the use of a time-reversible Markov model when this is
process by increasing the value of $K$.

Table 7. BIC values for different HAL-HAS models

<table>
<thead>
<tr>
<th>Model</th>
<th>$K$</th>
<th>$FP^a$</th>
<th>log $L$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAL-HAS1</td>
<td>2</td>
<td>103</td>
<td>263 116.24</td>
<td>130 958.74</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>153</td>
<td>263 347.46</td>
<td>130 958.74</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>203</td>
<td>263 911.93</td>
<td>130 958.74</td>
</tr>
<tr>
<td>HAL-HAS2</td>
<td>2</td>
<td>83</td>
<td>262 966.44</td>
<td>130 104.10</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>113</td>
<td>263 212.70</td>
<td>130 104.43</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>143</td>
<td>263 333.53</td>
<td>130 905.04</td>
</tr>
<tr>
<td>HAL-HAS3</td>
<td>2</td>
<td>71</td>
<td>262 979.90</td>
<td>131 110.25</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>89</td>
<td>263 122.36</td>
<td>130 107.20</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>103</td>
<td>263 281.14</td>
<td>130 090.61</td>
</tr>
<tr>
<td>HAL-HAS4</td>
<td>2</td>
<td>68</td>
<td>262 996.79</td>
<td>131 136.18</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>83</td>
<td>263 059.92</td>
<td>130 107.84</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>98</td>
<td>263 177.09</td>
<td>130 066.53</td>
</tr>
<tr>
<td>HAL-HAS5</td>
<td>2</td>
<td>55</td>
<td>263 120.02</td>
<td>131 267.04</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>57</td>
<td>263 136.06</td>
<td>131 264.40</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>59</td>
<td>263 152.13</td>
<td>131 261.78</td>
</tr>
</tbody>
</table>

$^a$Number of free parameters.

Table 8. BIC values for GTR+I+\ model

<table>
<thead>
<tr>
<th>$K$</th>
<th>log $L$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>264 420.2</td>
<td>132 003.50</td>
</tr>
<tr>
<td>3</td>
<td>264 362.4</td>
<td>132 006.40</td>
</tr>
<tr>
<td>4</td>
<td>264 252.0</td>
<td>132 006.80</td>
</tr>
<tr>
<td>5</td>
<td>264 302.8</td>
<td>132 006.80</td>
</tr>
<tr>
<td>6</td>
<td>264 362.4</td>
<td>132 006.80</td>
</tr>
<tr>
<td>7</td>
<td>264 252.0</td>
<td>132 006.80</td>
</tr>
<tr>
<td>8</td>
<td>264 257.8</td>
<td>132 006.80</td>
</tr>
<tr>
<td>9</td>
<td>264 252.0</td>
<td>132 006.80</td>
</tr>
<tr>
<td>10</td>
<td>264 252.0</td>
<td>132 006.80</td>
</tr>
<tr>
<td>11</td>
<td>264 252.0</td>
<td>132 006.80</td>
</tr>
<tr>
<td>12</td>
<td>264 252.0</td>
<td>132 006.80</td>
</tr>
</tbody>
</table>

Table 9. Inferred nucleotide frequencies for the variable sites

<table>
<thead>
<tr>
<th>Type</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_{\text{inv}}$</td>
<td>0.28072</td>
<td>0.16720</td>
<td>0.14621</td>
<td>0.40587</td>
</tr>
<tr>
<td>$\pi_1$</td>
<td>0.30998</td>
<td>0.46222</td>
<td>0.07136</td>
<td>0.20244</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>0.28978</td>
<td>0.30670</td>
<td>0.21582</td>
<td>0.18770</td>
</tr>
<tr>
<td>$\pi_3$</td>
<td>0.22566</td>
<td>0.36664</td>
<td>0.24075</td>
<td>0.16533</td>
</tr>
<tr>
<td>$\pi_4$</td>
<td>0.26943</td>
<td>0.29576</td>
<td>0.19049</td>
<td>0.25273</td>
</tr>
<tr>
<td>$\pi_5$</td>
<td>0.22716</td>
<td>0.25745</td>
<td>0.36576</td>
<td>0.14963</td>
</tr>
<tr>
<td>$\pi_6$</td>
<td>0.48104</td>
<td>0.04733</td>
<td>0.16779</td>
<td>0.30384</td>
</tr>
<tr>
<td>$\pi_7$</td>
<td>0.52131</td>
<td>0.11902</td>
<td>0.18221</td>
<td>0.17746</td>
</tr>
<tr>
<td>$\pi_8$</td>
<td>0.65712</td>
<td>0.13254</td>
<td>0.09685</td>
<td>0.11349</td>
</tr>
<tr>
<td>$\pi_9$</td>
<td>0.61030</td>
<td>0.18647</td>
<td>0.12977</td>
<td>0.07366</td>
</tr>
<tr>
<td>$\pi_{10}$</td>
<td>0.83758</td>
<td>0.06323</td>
<td>0.05000</td>
<td>0.08919</td>
</tr>
</tbody>
</table>

BIC values for join models, for both types of models,
obtained using these scenarios as the “free” models and
join models, respectively. For both types of models,
we implemented the HAS model using the discretized $I$ distribution,
and varied $K$ between 2 and 12. For the free model,
the optimal number of clusters was 4, and the BIC
value varied between 263 746 ($K = 2$) and 263 135 ($K = 12$).
For the join model, the optimal number of clusters was 4, and the BIC
value varied between 263 895 ($K = 2$) and 263 289 ($K = 12$). Hence,
the BIC values for the free and join models were higher than that for $H_{opt}$ (i.e., 262 966: Table 7).
Although the number of clusters and, thus,
the number of distinct rate matrices corresponding to the
HAL model, was the same for the optimal free model and $H_{opt}$,
the rate matrix arrangements differed at edge $e_2$ (see Fig. 7
for the distribution of rate matrices in $H_{opt}$),
which was assigned $R_3$ in the optimal free model.

Insights into the evolution of yeast genomes

Based on the analysis presented above, we were able to
assemble a more detailed picture of the evolution of
a substantial subset of sites from the yeast genome. The
ancestral sequence appears to have contained invariable
sites and two categories of variable sites, $V_1$ and $V_2$. The
inferred proportion of sites in these three categories of
sites is $0.49062$ for invariable sites, $0.39642$ for $V_1$, and 0.11295 for $V_2$.
The inferred nucleotide composition at the root
for the three sets of sites varied (Table 9),
with the largest difference in the proportion of C
(0.28978 vs. 0.41622 for $V_1$) and the smallest difference in the proportion of A
(0.28072 for $\pi_{\text{inv}}$ vs. 0.40587 for $\pi_1$).

After the divergence at node $I_0$ (Fig. 7), the variable
sites of the two descendant sequences evolved under
conditions approximated by $R_1$. A turn of events
then took place: following the divergence at node $I_1$,
the two descendant sequences evolved under different
conditions—the evolutionary process along edge $e_3$ was
approximated by $R_3$ while that along edge $e_4$ was
approximated by $R_2$. After the divergence at node $I_2$,
the two descendant sequences continued to evolve
under conditions approximated by $R_2$. A second turn of events then took place: after the divergence at node $I_3$, all descendant sequences evolved under conditions approximated by $R_1$. These results are consistent with those obtained during the preliminary survey of the yeast data (Figs. 5 and 6), highlighting the merits of using the matched-pairs tests of homogeneity as a way to obtain a preliminary insight into the conditions under which sequences might have evolved.

However, the evolutionary process at the variable sites is more complex, because the optimal model of evolution for the yeast data is a HAL-HAS2 model, so there were two sets of $\pi$-vectors (one each for $V_1$ and $V_2$) but only one set of $S$ (one for both $V_1$ and $V_2$). Furthermore, the proportions of A, C, G, and T varied from one $\pi$-vector to another (Table 9), and $S$ varied from one group of edges to another (Table 10). This illustrates a complexity of the evolutionary process that has not been realized previously for these data.

Table 10. Inferred values of $S$ for the yeast data

<table>
<thead>
<tr>
<th>Type</th>
<th>$S_{12}$</th>
<th>$S_{13}$</th>
<th>$S_{14}$</th>
<th>$S_{23}$</th>
<th>$S_{24}$</th>
<th>$S_{34}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s^1$</td>
<td>0.56567</td>
<td>2.36542</td>
<td>0.44573</td>
<td>1.54208</td>
<td>2.62548</td>
<td>0.25079</td>
</tr>
<tr>
<td>$s^2$</td>
<td>1.21057</td>
<td>1.99827</td>
<td>0.74843</td>
<td>2.25264</td>
<td>1.47400</td>
<td>0.49189</td>
</tr>
<tr>
<td>$s^3$</td>
<td>1.09277</td>
<td>1.48509</td>
<td>1.00938</td>
<td>1.80375</td>
<td>2.18375</td>
<td>0.44853</td>
</tr>
<tr>
<td>$s^4$</td>
<td>2.22360</td>
<td>1.98932</td>
<td>1.12828</td>
<td>5.84010</td>
<td>2.98071</td>
<td>0.68544</td>
</tr>
</tbody>
</table>

Figure 8 shows the phylogenetic trees with edge lengths drawn to scale for the two sets of variable sites. As the edges leading to Calb in the two trees were similar in length, we concluded that substitutions accumulating along the edges leading to the other taxa must have occurred at a higher rate, with the rate for sites belonging to $V_1$ being lower than that for sites belonging to $V_2$. Another interesting difference between the two trees in Figure 8 was that the accumulation of substitutions for sites belonging to $V_1$ dropped to almost zero along the lineages leading to $S_{bay}$, $Skud$, $Smik$, $Spar$, and $Scer$. By contrast, the accumulation of substitutions for sites belonging to $V_2$ remained well above zero and appeared fairly constant since the divergence of the lineage leading to $Skud$.

Although we cannot rule out that stochastic error and/or systematic error (due to model misspecification) may be factors contributing to the differences between the two trees, we note that a possible reason for the lower number of accumulated substitutions for sites belonging to $V_1$ might be natural selection. If this were the case, which seems reasonable because the data comprise second codon sites, then what Figure 8a suggests is a gradual increase in the selective pressure on sites belonging to $V_1$; this increase gradually caused the sites to stop evolving altogether along the lineages leading to $S_{bay}$, $Skud$, $Smik$, $Spar$, and $Scer$. In other words, what might be seen here is a shift in a set of sites from being variable to being invariable in five of the eight taxa. Although we cannot rule out that stochastic error and/or systematic error (due to model misspecification) may be factors contributing to the differences between the two trees, we note that a possible reason for the lower number of accumulated substitutions for sites belonging to $V_1$ might be natural selection. If this were the case, which seems reasonable because the data comprise second codon sites, then what Figure 8a suggests is a gradual increase in the selective pressure on sites belonging to $V_1$; this increase gradually caused the sites to stop evolving altogether along the lineages leading to $S_{bay}$, $Skud$, $Smik$, $Spar$, and $Scer$. In other words, what might be seen here is a shift in a set of sites from being variable to being invariable in five of the eight taxa.

An ability to accurately assign individual sites to specific rate categories might be particularly beneficial for data generated by concatenating multiple genes, such as the yeast data. For example, an investigator might...
FIGURE 9. Distribution of $\Delta \log L = \log L_{V_1} - \log L_{V_2}$ along the concatenated alignments of the 106 yeast genes.

Like to determine whether the rate categories are directly associated with the individual genes in the alignment. Finding this to be the case would support a gene-specific, model-selection approach developed by Lanfear et al. (2012); a finding to the contrary would support the development of another approach.

Given that our algorithm for identifying the optimal HAS model allows us to determine what rate category each variable sites is most likely to belong to, it became relevant to determine whether sites belonging to $V_1$ or $V_2$ form distinct clusters, with runs of sites belonging to $V_1$, followed by runs of sites belonging to $V_2$, and so forth.

For the yeast data, we found 12,801 sites (out of 42,337 second codon sites) to be variant sites (i.e., variable sites with differences across the sequences). For each of these sites, we calculated $\Delta \log L$ and charted it against the position in the alignment (Fig. 9). If $\Delta \log L > 0$ for a given site, then that site most likely belongs to $V_1$; otherwise, the site most likely belongs to $V_2$. Given this criterion and the results in Figure 9, there is very little visual evidence in support of sites belonging to either $V_1$ or $V_2$ clustering contiguously across the alignment of the 106 genes. In other words, our analysis lends little support for the gene-specific assignment of rate matrices currently being used in phylogenomic research (e.g., Lanfear et al. 2012; Leavitt et al. 2013). On the contrary, our study advocates using fewer rate categories over the sites (and more rate matrices over the tree)—this strategy, however, needs to be tested further through studies of other data sets.

CONCLUSION

We introduce two new algorithms for identifying the optimal and near-optimal HAL-HAS models of evolution for a given tree topology. Given that many phylogenetic data sets show evidence of HAL and/or HAS, and the number of possible HAL-HAS models increases super-exponentially with a growth in the number of sequences, our algorithms are clearly desirable.

Our algorithms complement a small number of algorithms that have been used to identify optimal and near-optimal HAL and HAL-HAS models (e.g., Blanquart and Lartillot 2008; Jayaswal et al. 2011a; Dutheil et al. 2012). In this context, it is worth noting that the choice of method should be data-driven: that is, an initial survey of the data is needed to determine the best analytical strategy. For example, if the number of taxa is small (say ≤7), then the CORE algorithm should be preferred over the SHAL algorithm for identifying the optimal HAL model. This is because the former algorithm needs to evaluate fewer models when identifying the optimal HAL model. By contrast, if the number of taxa is large, then the SHAL algorithm should be preferred because, for a larger number of taxa, some of the models evaluated by the CORE algorithm might not be identifiable. For a given HAL model, the optimal HAS model can be identified by comparing results obtained using our family of HAS models to those obtained using the commonly used $\Gamma$-distributed HAS model, as implemented by Dutheil et al. (2012).

Our family of HAS models makes it possible to evaluate multiple sources of HAS and identify the most likely one. One of our HAS models, namely, GTR+$\Gamma$, is similar to the non-$\Gamma$ HAS model implemented in release 3.1 of PhyML (Guindon et al. 2010). Although both the models estimate $\omega_1, \cdots, \omega_K$, the number of free parameters differs by one. For $K$ rate categories, our model assumes that the first-edge length multiplier ($\omega_1$) is one and the remaining scalar multipliers ($\omega_2, \cdots, \omega_K$) are estimated. In contrast, the non-$\Gamma$ model estimates all the scalar multipliers.

One limitation of our algorithm is that all the sites, irrespective of which rate category they belong to, have the same HAL model. As this assumption may not hold true for real data, more complex algorithms are needed to obtain even better HAL-HAS models of evolution.
Another limitation of our algorithm is that no attempt is made to reduce the number of free parameters in $S$ and $\pi$. For example, Table 10 suggests that it might be possible to equate: (a) $S_{13}$ and $S_{23}$ in $S^1$; (b) $S_{13}$ and $S_{23}$ in $S^2$; (c) $S_{12}$ and $S_{14}$ in $S^3$; and (d) $S_{12}$ and $S_{13}$ in $S^4$ (determined by obtaining all pairwise ratios of elements in $S$ and identifying the pair with the smallest ratio $\geq 1$). This reduction might lead to more accurate and precise parameter estimates.

A further limitation of our algorithm is that it assumes that every site has evolved on the same tree topology. This assumption might not be appropriate if, for example, some of the genes are paralogous genes or have undergone recombination or gene conversion.

A final limitation of our algorithm is that it does not enable a search of tree space. However, in this regard, our algorithm is similar to several other methods for identifying optimal HAL models (e.g., Jayaswal et al. 2011; Dutheil et al. 2012; Goussin et al. 2013). As a change in tree topology may alter what is inferred to be the optimal HAL-HAS model, a two-step process that alternates between searching for the best tree and searching for the best HAL-HAS model is desirable. The former search could be done using algorithms such as nearest-neighbor interchange, subtree pruning and regrafting, and tree bisection and reconnection (Felsenstein 2004).

A HAL or HAS model (or a combination thereof) can have multiple maxima, implying that parameter estimation should be done repeatedly with different starting values (for an example, see Appendix 2). This can significantly increase the overall execution time for a HAL or HAS model compared to the time-reversible family of models, thus making the former models less attractive. However, a comparison of the HAL-HAS models with the family of time-reversible models is a must if identifying the optimal model of evolution is the objective, as demonstrated using the yeast data set.

Finally, our HAL-HAS model of evolution may be used to identify edges over which sites change from being variable and to being invariable. For example, our analysis of the yeast data suggested that a large proportion of the variable sites (i.e., 39.6% of all sites in the alignment) have an extremely low rate of evolution (close to zero) over the subtree comprising $Shay$, $Skud$, $Smk$, $Scr$, and $Spar$.

**SUPPLEMENTARY MATERIAL**

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.9h6f8.

**ACKNOWLEDGMENTS**

We thank A. Rokas for sending the yeast genome data, S. Kalyaanamoorthy, O. Edwards, and M. Woodhams for their constructive suggestions on the manuscript, and F.E. Anderson, E. Susko, B. Boussau, K. Schliep, and an anonymous reviewer for their many constructive comments, which helped to improve the manuscript further. VJ was supported by a Postdoctoral Fellowship from the Australian Research Council, and TKFW was supported by a Postdoctoral Fellowship from OCE Science Leader funds awarded to LSJ by CSIRO.

**APPENDIX 1**

(a) Initial Parameter Values for SHAL and CORE Algorithms

To obtain the optimal HAL model using the SHAL or CORE algorithm, we set

1. $f_0 = [0.25 \quad 0.25 \quad 0.25]^T$
2. $\pi = [0.25 \quad 0.25 \quad 0.25]^T$ for all edges
3. $S_{12} = \cdots = S_{14} = 1$ for all edges
4. $t_1 = \cdots = t_{28} = 1$
5. $\beta$ was set to 75% of constant sites
6. $\pi_{inv} = [0.25 \quad 0.25 \quad 0.25]^T$

(b) Initial Parameter Values for HAS Models

To obtain the optimal HAS model for a given HAL model, we set

1. $f_0 = [0.25 \quad 0.25 \quad 0.25]^T$ for all rate categories
2. $\pi = [0.25 \quad 0.25 \quad 0.25]^T$ for all edges and rate categories
3. $S_{12} = \cdots = S_{14} = 1$ for all edges and rate categories
4. $t_1 = \cdots = t_{28} = 1$ for all rate categories
5. $\beta$ was set to 75% of constant sites and $\alpha_1 = \cdots = \alpha_K$, where $K$ denotes the number of categories for variable sites
6. $\pi_{inv} = [0.25 \quad 0.25 \quad 0.25]^T$

(c) Parameter Values for Data Generated Under the GTR+$\Gamma$ Model

To obtain simulated data under the GTR+$\Gamma_2$ model, we used Seq-Gen (Rambaut and Grassly 1997), the tree shown in Figure 2b, and the parameter values shown below:

1. $\pi_1 = 0.333732$, $\pi_2 = 0.214585$, $\pi_3 = 0.073465$, and $\pi_4 = 0.378218$
2. $S_{12} = 1.222060$, $S_{13} = 6.496830$, $S_{14} = 0.473932$, $S_{23} = 6.060950$, $S_{34} = 0.971132$, and $S_{44} = 0.684187$
3. Discretized $\Gamma$ distribution, with $K=2$ rate categories, and a shape parameter of 1.0
To obtain simulated data under the GTR+Y model, we used Hetero Version 2.0, the tree shown in Figure 2a, and the parameter values shown below:

1. $f_0 = \pi$ and $\pi_1 = 0.333732$, $\pi_2 = 0.214585$, $\pi_3 = 0.073465$, and $\pi_4 = 0.378218$ for both categories
2. $S_{12} = 1.22060$, $S_{13} = 6.496830$, $S_{14} = 0.473932$, $S_{23} = 6.060950$, $S_{24} = 0.971132$, and $S_{34} = 0.684187$ for both rate categories and for all edges
3. $\beta = 0.000000$, $a_1 = 0.327012$ and $a_2 = 0.672988$
4. $t_1 = \cdots = t_{28} = 2$ for the first rate category
5. $p_2 = 0.053969$ for the second rate category

We note that the GTR+Y model assumes SRH and, therefore, uses the constraint $f_0 = \pi$ during data generation and parameter estimation.

**APPENDIX 2**

We demonstrate the occurrence of multiple maxima using two models, the first model considers only HAL and the second model considers only HAS.

**HAL-only Model**

We analyzed the eight-taxon yeast data using the HAL model specified in Figure 7. We considered 100 sets of initial parameter values and these were generated as follows:

1. $f_0$: generated four random numbers from a uniform distribution in the interval $[0.1, 1]$ and scaled the numbers so that they summed to 1.
2. $\pi$: generated four vectors of equilibrium frequencies, one each for the four distinct rate matrices in the HAL model, using the same procedure as above.
3. $S = \{S_{12}, \ldots, S_{34}\}$: generated six random numbers from a uniform distribution in the interval $[0.1, 1]$. Next, we assigned the same $S$ to all the edges.
4. $t_1, \ldots, t_{28}, t_2$: generated $28 - 2$ random numbers from a uniform distribution in the interval $[0.1, 1]$
5. $\beta$: generated a random number from a uniform distribution in the interval $[0.1, 1]$
6. $\pi^{(0)}$, we generated $\pi^{(0)}$ using the same procedure as that for obtaining the vector of equilibrium frequencies.

For the 100 different initial parameter values, Table A1 shows the log-likelihoods at convergence and the proportion of times a given value occurred.

**TABLE A1.** Occurrence of multiple maxima

<table>
<thead>
<tr>
<th>Log-likelihood</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>−131570.71</td>
<td>0.01</td>
</tr>
<tr>
<td>−131570.72</td>
<td>0.03</td>
</tr>
<tr>
<td>−131570.73</td>
<td>0.10</td>
</tr>
<tr>
<td>−131570.74</td>
<td>0.08</td>
</tr>
<tr>
<td>−131570.75</td>
<td>0.21</td>
</tr>
<tr>
<td>−131570.76</td>
<td>0.45</td>
</tr>
<tr>
<td>−131570.77</td>
<td>0.03</td>
</tr>
<tr>
<td>−131582.84</td>
<td>0.01</td>
</tr>
<tr>
<td>−131582.89</td>
<td>0.01</td>
</tr>
<tr>
<td>−131584.37</td>
<td>0.01</td>
</tr>
<tr>
<td>−131584.82</td>
<td>0.01</td>
</tr>
<tr>
<td>−132638.86</td>
<td>0.01</td>
</tr>
<tr>
<td>−132654.63</td>
<td>0.01</td>
</tr>
<tr>
<td>−132657.92</td>
<td>0.01</td>
</tr>
<tr>
<td>−132657.99</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**HAS-only Model**

To demonstrate the occurrence of multiple maxima in a model that considers HAS but no HAL, we analyzed one of the pseudo-data sets generated under the GTR+Y model (refer subsection “Comparison of HAS models that consider scalar multiples of edge lengths”) using the same initial parameter values as specified in Appendix 1(d), except for $p_2$. We considered three values of $p_2$, namely, 0.5, 1, and 10. The log-likelihoods obtained using the three $p_2$ values were $−132706.61$, $−132707.64$, and $−132707.26$, respectively. Therefore, even for models with no HAS, changing the initial value for just one of the parameters results in slightly different log-likelihoods.

**REFERENCES**


