Phylogenetic Substitution Models for Detecting Heterotachy during Plastid Evolution

Simon Whelan,*1 Benjamin P. Blackburne,1 and Matthew Spencer2

1Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom
2School of Environmental Sciences, University of Liverpool, Liverpool, United Kingdom

*Corresponding author: E-mail: simon.whelan@manchester.ac.uk.
Associate editor: Barbara Holland

Abstract

There is widespread evidence of lineage-specific rate variation, known as heterotachy, during protein evolution. Changes in the structural and functional constraints acting on a protein can lead to heterotachy, and it is plausible that such changes, known as covarion shifts, may affect many amino acids at once. Several previous attempts to model heterotachy have used covarion models, where the sequence undergoes covarion drift, whereby each site may switch independently among a set of discrete classes having different substitution rates. However, such independent switching may not capture biologically important events where the selective forces acting on a protein affect many sites at once. We describe a new class of models that allow the rates of substitution and switching to vary among branches of a phylogenetic tree. Such models are better able to handle covarion shifts. We apply these models to a set of genes occurring in nonphotosynthetic bacteria, cyanobacteria, and the plastids of green and red algae. We find that 4/5 genes show evidence of some form of rate switching and that 3/5 genes show evidence that the relative switching rate differs among taxonomic groups. We conclude that covarion shifts may be frequent during the deep evolution of plastid genes and that our methodology may provide a powerful new tool for investigating such shifts in other systems.

Key words: phylogenetics, substitution models, heterotachy, plastid evolution.

Introduction

A key aspect of many evolutionary studies is the use of genomic sequence data to infer a phylogenetic tree that describes the ancestral relationships between species. For example, these trees have provided insights into important events in early evolution (Baldauf et al. 2000; Dunn et al. 2008; Inagaki et al. 2009), the phylogeographic history of organisms (Torroni et al. 2006), and the interactions between pathogens and hosts (Rambaut et al. 2004). Maximum likelihood (ML) and Bayesian inference are now the dominant forms of phylogenetic inference. The performance of these methods is dependent on a substitution model, a stochastic model used to describe the relative frequency of substitutions between sequence characters, such as nucleotides or amino acids. When the substitution model adequately reflects the “true” evolutionary processes, the methods perform well and have appealing statistical properties, including consistency (Rogers 1997) and robustness (Delsuc et al. 2005; Whelan 2008a). If the substitution model is misspecified (poorly describes the true process), inferences can become systematically biased and may converge on the wrong answer with increasing probability as the amount of data increases.

One area of model misspecification that has attracted a lot of attention recently is lineage-specific rate variation (Inagaki et al. 2004; Kolaczkowski and Thornton 2004, 2008; Whelan 2008b), where the rate of evolution of a site in a protein changes over the course of evolution. This phenomenon is frequently referred to as heterotachy and has been found in many different genes across many different species, including RNA sequences in eukaryotes (Philippe and Germot 2000) and elongation factor 1α in microsporidia (Inagaki et al. 2004). Empirical studies suggest that heterotachy makes accurate tree inference difficult, causing statistical methods of tree inference to suffer from long-branch attraction (LBA) artifacts (Philippe et al. 2005), a problem more usually associated with maximum parsimony inference (Felsenstein 1978). These findings have been supported by simulation studies, which have shown that LBA can be caused by a variety of forms of lineage-specific rate variation, including changes in among-site rate variation (Kolaczkowski and Thornton 2004), inadequate descriptions of codon evolution (Whelan 2008c), and more complex patterns of rate changes. Under some conditions, sequence alignments formed by mixing data generated by two different sets of branch lengths on a single-tree topology can result in several different tree topologies having precisely the same ML, which means that the tree parameter is not identifiable (Matsen and Steel 2007).

One approach for dealing with lineage-specific rate variation is to incorporate it into the substitution model. Several modeling approaches have been developed, describing it in two distinct manners. The first approach, originally used for simulation by Kolaczkowski and Thornton (2008), describes substitution as a mixture model of different Markov processes, each with its own distinct set of branch lengths. Their study suggests that mixture models could be used to accurately estimate phylogenetic trees under a variety of different forms of heterotachy. Similar
approaches have also been used to study how a change in the proportion of variable sites on a lineage affects tree inference (Gruenheit et al. 2008). A second popular approach, and the focus of this paper, uses generalizations of the covarion model of Tuffley and Steel (1998), which are referred to as covariation-type models or temporal hidden Markov models (THMMs) (Whelan 2008b). These models describe evolution through a series of substitution models, representing the hidden states in the model, and a model of switching between these hidden states, where the rates of change between states are analogous to the transition rates in a standard hidden Markov model. This approach allows individual sites to change how they evolve during time and have the potential to capture biologically important information. However, in the majority of existing studies, the rate of switching is constant through time, resulting in sites gradually changing their substitution processes over time at a rate directly proportional to the substitution rate; a process referred to as covarion drift (e.g., Ané et al. 2005). These models fail to capture covarion shifts, where there are rapid and correlated shifts in substitution process across many sites, which may be independent of substitution rate. Covarion shifts can be caused by a range of possible scenarios, including changes in selective pressures caused by movement into a different environment, a change in lifestyle (e.g., parasitism), long-term changes in gene expression, and/or a shift in protein structure.

One example where lineage-specific rate variation has been carefully examined is during the endosymbiosis event leading to the origins of plastids (Lockhart et al. 2006). Attempts to resolve the evolutionary relationships between several classes of plastids and their relatives have proved troublesome, with significant phylogenetic incongruence occurring between genes inferred from RNA polymerase subunits and other plastid genes (Martin et al. 1998). Some of these difficulties have been attributed to covarion shifts during plastid evolution, in particular, changes in the number of highly constrained or invariant sites in proteins. Lockhart et al. (2006) demonstrated that apparent lineage-specific rate variation in different plastid groups and their relatives was the result of changes in the relative numbers of highly constrained sites in the proteins. These covarion shifts may be the cause of error in phylogenetic tree inference (Lockhart and Steel 2005). Several other studies have also provided detailed examinations of heterotachy, both as covarion drift and covarion shift in a single system (e.g., Gaucher et al. 2001; Inagaki et al. 2004; Ané et al. 2005), but there are currently few general methods available for identifying and characterizing heterotachy.

This study presents a new substitution model that builds upon current “covariation-style” THMMs in a manner that can capture covarion shifts. Our model describes rate variation as random draws from a discrete Ω-distribution and has an additional hidden class describing invariant sites. Rather than having a constant rate of switching between hidden classes, we decouple substitution rate and switching rate, which allows each to have its own rate on each branch in a tree. We apply our model to the plastid data of Lockhart et al. (2006) and, using likelihood ratio tests (LRTs), find further evidence for covarion shift. We also demonstrate that our model can provide a good visualization of where shifts occur on a tree. We use simulations to show that our model performs as expected and that our findings are unlikely to be artifacts. Furthermore, we show that the asymptotic distributions usually used in LRTs tend to be conservative for our model and that parametric bootstrap tests are more powerful.

Materials and Methods

Sequence Data and Phylogenetic Trees

The sequences used in this study are from Lockhart et al. (2006) and were downloaded from the Supplementary Material online (http://mbe.oxfordjournals.org/cgi/content/full/msj005/D1). Each of the 5 alignments contains sequences from 16 species: 4 nonphotosynthetic bacteria, 4 photosynthetic cyanobacteria, 4 green algae, and 4 red algae. Each alignment contains one or more genes found in plastids and their homologues from free-living organisms. Specifically, they contain 1) ATPase α-subunit gene (AtpA, with length 476 amino acids); 2) ATPase β-subunit gene (AtpB, 424 amino acids); 3) RNA polymerase subunit B (RpoB, 695 amino acids); 4) concatenated sequences from RNA polymerase subunits C1 and C2 (RpoC1C2, 832 amino acids); and 5) protein synthesis elongation factor Tu (TufA, 235 amino acids). These alignments were concatenated and then used to estimate a phylogenetic tree that matches the broad relationships between the groups of taxa, with the red and green algae grouping together. This tree topology, shown in figure 1, was taken as fixed for all subsequent analyses, including simulations. All visualization of phylogenetic trees was performed using Dendroscope (Huson et al. 2007).

Substitution Models and Temporal Heterogeneity

All the amino acid substitution models examined in this study are special cases of a general branch-specific THMM, which incorporates temporal heterogeneity in rate (see Whelan 2008b). Its relationship to other models, such as those of Galtier (2001) and Huelsenbeck (2002), can be assessed by comparison to the general covarion model of Wang et al. (2007). Specifically, on a single branch, our model generalizes the Galtier model by adding an extra state with a zero substitution rate. (Note, however, that the resulting model is neither a restriction nor a generalization of the Huelsenbeck or Wang models.) The elements of the instantaneous rate matrix on branch \( b \) between observable characters \( i \) and \( j \) and between hidden states \( k \) and \( l \) are defined as

\[
Q(b)_{ij}^k = \pi^l \pi_j^l \begin{cases} 
- \sigma(b) & \text{for } i \neq j; k = l, \\
0 & \text{for } i = j; k \neq l, \\
\rho_k & \text{otherwise.} 
\end{cases}
\]

The stationary probability of hidden state \( l \) and observable character \( j \) is \( \pi^l_j = \pi^l \pi_j^l \), where \( \pi^l \) is the stationary probability of hidden state \( l \).
probability of hidden state \( l \) and \( \pi_j \) is the stationary probability of observable character \( j \). The relative substitution rates of individual hidden classes are defined by \( r^k \) and the exchangeability parameters, \( S_{ij} \), describe the relative rates of change between different amino acids. The rate of transition between all hidden states on branch \( b \) is controlled by the switching parameter \( r(b) \). The model is scaled so that the mean rate of substitution, \( l_{\text{subs}} = \sum_{k=1}^{5} \pi_k Q_{ij}^k \), is equal to 1, which means that the branch lengths on trees can be interpreted in the usual manner as the expected number of substitutions per site, and a unit of evolutionary time is defined as one expected substitution per site. For each branch, the mean rate of switching can therefore be calculated as \( l_{\text{switch}} = \sum_{k \neq j} \pi_k Q_{ij}^k \), and the value of \( l_{\text{switch}} \) estimated from data represents the expected number of switches per site per unit of evolutionary time. This model structure decouples the rate of substitution and switching occurring on a branch enabling comparison of the rates of substitution and switching.

The most general model we use has five hidden states, with the first hidden state representing invariable sites, \( r^1 = 0 \), and occurring with probability \( \pi^1 = p_{iv} \). The remaining four states occur with equal probability, \( \pi^2 = \pi^3 = \pi^4 = \pi^5 = (1 - p_{iv})/4 \), and have their rates defined as the discretized rates of a \( \Gamma \)-distribution, controlled in the usual manner by the parameter \( \alpha \) (Yang 1994). The \( S_{ij} \) values are taken from the Whelan and Goldman (WAG) model (Whelan and Goldman 2001) and the amino acid frequencies, \( p_j \), are taken as the relative frequency with which amino acids occur in the sequence data. We examined our data using several different amino acid substitution models, and the choice of model had little impact on the inference (data not shown). In the remainder of this paper, we refer to this model as WAG + F + bCov(I,\( I' \)), where bCov(...) refers to the branch-specific covarion-type model.

The remaining models we investigate can be formed by placing restrictions on WAG + F + bCov(I,\( I' \)). The group-specific covarion-type model, WAG + F + gCov(I,\( I' \)), is formed by restricting the lineages within each of the five groups depicted in figure 1 to have the same switching rate. The four different groups containing lineages leading to extant organisms each have their own switching rates, defined as \( r(\text{bact}) \) for the nonphotosynthetic bacteria, \( r(\text{cyan}) \) for the cyanobacteria, \( r(\text{green}) \) for the green algae, and \( r(\text{red}) \) for the red algae. The five ancestral branches linking the four extant groups also has its own rate of switching, \( r(\text{anc}) \).

Placing the restriction that all branch-specific switching rates are equal, \( \sigma = \sigma(b_1) = \sigma(b_2) = \ldots = \sigma(b_{2n-3}) \), results in a model of covarion drift, WAG + F + Cov(I,\( I' \)). This model is comparable to existing THMMs that allow switching between hidden states that vary by rate, with the rate of switching on any given branch being proportional to the length of the branch. The standard WAG + F + I + \( I' \) model, which describes rate heterogeneity across sites, but without changes in relative rates over time, can then be formed by setting \( \sigma \) to zero. Table 1

**Fig. 1** The tree used in this study to describe the relationship between the 16 bacteria and plastids. Dotted lines are used to identify the branches used to group the lineages in the tree into five separate groups consisting of the nonphotosynthetic bacteria (bact), the cyanobacteria (cyan), the green algae (green), and the red algae (red). Note branch lengths are not drawn to scale.
Whelan et al. · doi:10.1093/molbev/msq215

Table 1. Relationship between the Models Presented.

<table>
<thead>
<tr>
<th>Model name</th>
<th>Parameters</th>
<th>Model df.</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAG + F</td>
<td>$\pi^a$</td>
<td>19</td>
</tr>
<tr>
<td>WAG + F + I + $I'$</td>
<td>$\pi, p_{inv}$</td>
<td>21</td>
</tr>
<tr>
<td>WAG + F + Cov($I'$)</td>
<td>$\pi, p_{inv}, \alpha^d$</td>
<td>22</td>
</tr>
<tr>
<td>WAG + F + gCov($I'$)</td>
<td>$\sigma(bac), \sigma(cyan), \sigma(green), \sigma(red)^e$</td>
<td>26</td>
</tr>
<tr>
<td>WAG + F + bCov($I'$)</td>
<td>$\pi, p_{inv}, \alpha, \sigma_{inv}$</td>
<td>50</td>
</tr>
</tbody>
</table>

NOTE.—df, degrees of freedom.

$^a$ The stationary probability of amino acids occurring.

$^b$ The proportion of invariable sites in the model.

$^c$ The parameter describing the variance of $I'$-distributed rates-across-sites.

$^d$ The THMM switching parameter.

$^e$ Group-specific switching parameters.

$^f$ Branch-specific switching parameters.

details the parameters associated with each model and its constituent degrees of freedom.

Implementation, Likelihood Computation, and Statistical Testing

All the models discussed above are implemented in the phylogenetic program LeaPhy (Whelan 2007), and likelihood computations are conducted in the usual manner (Felsenstein 2003). Parameters are estimated using standard numerical optimization techniques, with several different starting positions investigated to ensure good optima are located.

Comparisons between models are conducted using LRTs whereby the log-likelihood difference between the null and alternative hypotheses is calculated as $\delta = \ln L(H_{alt}) - \ln L(H_{null})$. All the LRTs conducted are between nested models, where the null hypothesis can be constructed by placing restrictions on the alternative hypothesis. Under standard regularity conditions, comparisons between nested models that differ by $n$ degrees of freedom can be conducted by comparing twice the log-likelihood difference, $2\delta$, to the appropriate critical values of a $\chi^2_n$ distribution.

Simulation Conditions for Investigating $p_{inv}$ and Null Distributions of $2\delta$

Simulation approaches are used to investigate the performance of the WAG + F + bCov($I'$) model. We first investigate whether changes in the proportion of invariable sites on the lineages leading to the four main groups are adequate to explain the patterns of switching we observe in the real data. We perform this simulation using lineage-specific_seq_gen (Shavit Grievink et al. 2008) under WAG + F + I, with the proportion of invariable sites, $p_{inv}$, differing between the bact, cyan, green, red, and anc groups. The first four values of $p_{inv}$ are estimated under WAG + F + I by ML from the four extant taxa for each group. The ancestral value of $p_{inv}$ for anc is taken to be that estimated from all 16 taxa.

The comparison between WAG + F + Cov($I'$) and WAG + F + bCov($I'$) requires a hypothesis test with 28 degrees of freedom. The number of parameters is large compared with the amount of sequence data available for inference, which is reflected in relatively low increases in likelihood for several data sets. We use a parametric bootstrap test to investigate whether the asymptotic $\chi^2_{28}$ null distribution is appropriate for these relatively small sample sizes. Using purpose written simulation software, we generate $R$ data sets of the same size as the real data under WAG + F + Cov($I'$), using ML estimates of parameters taken from the real data. For each replicate data set, $r$, both models are fitted and a bootstrap likelihood ratio statistic $2\delta^*$ is obtained. To assess statistical significance of the values of $2\delta$ observed in our real data, we compare it with the 95% mark of the bootstrapped distribution of $2\delta^*$ (Davison and Hinkley 1997).

Results

Modeling and Parameter Estimation in Plastid Genes

We apply a range of substitution models (see Table 1) to the plastid data taken from Lockhart et al. (2006) to investigate temporal heterogeneity in evolutionary rate. Table 2 shows the log-likelihoods of the models for each gene and a selection of ML estimates of the parameter associated with them. The log-likelihood of the model, ln$L$, clearly increases in all genes as the model becomes more complex, matching expectations from the structure of the models, although which models provide a significant improvement in likelihood varies between genes (Testing significance of temporal heterogeneity).

The ML estimates of parameter values show consistent patterns for most genes, suggesting that the models are performing in a predictable and coherent manner (Table 2). The proportion of invariable sites, $p_{inv}$, tends to increase substantially when temporal heterogeneity is incorporated into the model through a THMM. For example, in AtpA, the value of $p_{inv}$ increases from 0.14 to 0.27 when moving between the WAG + F + I + $I'$ and the WAG + F + Cov($I'$) models. In most model comparisons, there is also a noticeable, but smaller, increase in $p_{inv}$ when adding to the model group or branch-specific rates of switching. In AtpB, the value of $p_{inv}$ increases from 0.24 under WAG + F + Cov($I'$) to 0.33 under both WAG + F + gCov($I'$) and WAG + F + bCov($I'$), respectively. The increase in the value of $p_{inv}$ also tends to be matched by an increase in the value of the rates-across-sites parameter $x$, representing a reduction in the variance of the $I'$-distribution. Our findings match previous observations that models of simple spatial heterogeneity, such as $I'$-distributed rates-across-sites, tend to compensate for simple forms of heterotachy (Wang et al. 2007; Whelan 2008b).

Inferred Patterns of Switching

The relative values of the substitution tree length and the switching tree length give an indication of how frequently rate switches occur. The amount of switching inferred to occur in all genes is relatively high and often increases when group or branch-specific switching is allowed. On RpoB, for
Table 2. Model likelihoods and parameter estimates.

<table>
<thead>
<tr>
<th>Gene</th>
<th>lnL</th>
<th>ΔlnL</th>
<th>$p_\text{err}$</th>
<th>$\alpha$</th>
<th>Subs length</th>
<th>Switch length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>~</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtpA</td>
<td>8127.06</td>
<td>8119.26</td>
<td>0.27</td>
<td>1.51</td>
<td>4.65</td>
<td>2.92</td>
</tr>
<tr>
<td></td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>RpoB</td>
<td>13079.93</td>
<td>13078.67</td>
<td>0.56</td>
<td>0.72</td>
<td>2.72</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>RpoC1C2</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td></td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>TufA</td>
<td>2899.52</td>
<td>2897.60</td>
<td>0.27</td>
<td>0.72</td>
<td>2.42</td>
<td>2.33</td>
</tr>
</tbody>
</table>

* Likelihood difference calculated between WAG + F + Cov(I,$\Gamma$) and WAG + F + I + $\Gamma$. Significance is tested by comparing twice the likelihood difference to the 95% mark of $\chi^2_2$ distribution, with ** used to label the rejection of the null hypothesis.

b Likelihood difference calculated between WAG + F + gCov(I,$\Gamma$) and WAG + F + Cov(I,$\Gamma$). Significance is tested by comparing twice the likelihood difference to the 95% mark of $\chi^2_2$ distribution, with * used to label the rejection of the null hypothesis.

c Likelihood difference calculated between WAG + F + bCov(I,$\Gamma$) and WAG + F + Cov(I,$\Gamma$). Significance is tested by comparing twice the likelihood difference to the 95% mark of $\chi^2_2$ distribution, with * used to label the rejection of the null hypothesis.

d Values describe the substitution tree length, which is the sum of the per branch estimates of the expected number of switches per site (changes between observable states).

e Values describe the switching tree length, which is the sum of the per branch estimates of the expected number of switches per site (changes between hidden states).

For example, the average rate of switching is (1.50/5.60 = 0.27) times the rate of substitution under WAG + F + Cov(I,$\Gamma$) and this increases to (2.46/5.58 = 0.44) under the branch-specific WAG + F + bCov(I,$\Gamma$) model. The TufA gene is an exception to this pattern, which may be because for this gene, there is little support (increase in lnL) for the models of temporal heterogeneity.

The first two columns in figure 2 show the shapes of the trees inferred under WAG + F + Cov(I,$\Gamma$). The “substitutions” tree shows what is typically thought of as a phylogenetic tree, with the branch lengths proportional to the expected number of substitutions per site whereas the “switches” tree has branch lengths proportional to the expected number of switches between hidden states per site. Note that the scaling factor for both trees is usually different, and the trees are presented at roughly the same size to allow easy visual inspection. The genes show the same patterns of substitution as those presented in Lockhart et al. (2006), with the Rpo genes showing long branches in the nonphotosynthetic bacterial and green algal lineages, whereas the other three genes tending to show accelerated rates only in the nonphotosynthetic bacterial lineages.

A similar, but more extreme, pattern is seen for the switching trees. In the Rpo genes, the nonphotosynthetic bacterial lineages and the green algal lineages both demonstrate significant switching, although there are a few lineages with low rates of switchings, such as Rickettsia in the RpoC1C2 gene. In contrast, the cyanobacterial and red algal lineages show typically low or zero rates of switching, with a few exceptions, such as the Guillardia lineage in RpoB. The AtpA and AtpB genes both show a strong tendency for high-switching rates on the nonphotosynthetic bacterial lineages and a tendency for low or zero rates of switching on the other lineages. The pattern of switching in TufA is similar to that of RpoB and RpoC1C2, but the pattern is less clear and there is more switching, particularly in the red algal lineages. This pattern is reflective of the lower support for WAG + F + Cov(I,$\Gamma$) relative to WAG + F + Cov(I,$\Gamma$) in TufA.

Testing Significance of Temporal Heterogeneity

The models presented in this study are nested versions of the general branch-specific model described in equation (1), enabling their comparison through likelihood ratio testing. The comparison between the null hypothesis of WAG + F + I + $\Gamma$ and the alternative hypothesis of WAG + F + Cov(I,$\Gamma$) tests whether there is significant evidence of gradual rate switching (covarion drift) during the evolution of a gene. The log-likelihood values, lnL, presented in table 2 show that in 4/5 of the genes examined WAG + F + Cov(I,$\Gamma$) provides a significantly better description of their evolution than WAG + F + I + $\Gamma$, providing evidence for temporal heterogeneity in plastid gene evolution. The only gene that does not provide a significant increase in likelihood in this comparison is TufA. There is less widespread evidence for group- and branch-specific variation in the patterns of switching (table 2). The group-specific WAG + F + gCov(I,$\Gamma$) model provides significant improvement over WAG + F + Cov(I,$\Gamma$) in AtpA, AtpB, and RpoC1C2. Using a conventional $\chi^2_4$ distribution, only RpoC1C2 provided evidence for branch-specific switching, with a significant increase in likelihood for the
comparison between WAG + F + Cov(I, I') and WAG + F + bCov(I, I').

The consistent pattern of switching shown in figure 2, and previous observations by Lockhart et al. (2006), suggests that branch-specific test may be conservative. We investigated this using a Monte Carlo simulation to empirically estimate the null distribution for comparing WAG + F + Cov(I, I') and WAG + F + bCov(I, I') in all genes. Figure 3 shows the cumulative density function (c.d.f.) for the simulated log-likelihood differences calculated between WAG + F + Cov(I, I') and WAG + F + bCov(I, I'), as well as the asymptotic $\chi^2_{28}$ distribution. The c.d.f.s under all simulation conditions are shifted to the left and their 95% marks indicate that in all cases, the asymptotic null distribution is excessively conservative for hypothesis testing. The branch-specific LRT in the real AtpA, for example, yields an improvement in log-likelihood of 17.86. This comparison is not significant when twice this value ($2\delta = 35.72$) is compared with 41.34, the 95% mark of the $\chi^2_{28}$ distribution. However, the value of $2\delta$ is significant when compared with the 95% mark of the $2\delta^*$ distribution produced from the parametric bootstrap. For AtpB, RpoB, and TufA, there is little evidence of covarion shifts, even with the parametric bootstrap test.

Investigating Switching Patterns Using Simulation

Figure 2 shows that the switching patterns observed in the genes tend to fall into two patterns. The RNA polymerase genes, RpoB and RpoC1C2, show rapid switching in the nonphotosynthetic bacterial and green algal lineages, whereas the AtpA, AtpB, and TufA genes show rapid switching occurring only in nonphotosynthetic bacterial lineages. We use a simulation approach to investigate whether our inferential approach could lead to structured artifacts or whether a shift in the proportion of invariable sites, $p_{inv}$, is adequate to describe the observed pattern.

The third and fourth columns in figure 2, headed simulated covarion drift, are produced from data simulated...
under the WAG + F + Cov(I, J) model using ML parameter estimates from the original data. Under this model, the rate of switching is directly proportional to the amount of substitution occurring in a branch (covarion drift). The shapes of the substitution and switching trees in columns 3 and 4 of figure 2 are those inferred from WAG + F + bCov(I, J), where the branch lengths shown are the median estimates from 100 simulations. For all families, the branch length estimates for the substitution trees (fig. 2, column 3) are comparable to those used for the simulation.

The pattern of switching (fig. 2, column 4) appears more complex than the pattern of substitutions and there is some variation between genes. For AtpA, AtpB, RpoB, and RpoC1C2, the patterns of switching are broadly comparable to the substitution patterns, although short branches and internal branches are frequently zero. For TufA, no internal branches appear to be nonzero and the branching pattern is unusual. We attribute these unusual branch-specific patterns of switching to the high variance of the estimator \( \sigma(b) \) caused by the fact that the expected number of switches occurring on a branch over the entire sequence is less than one. From our simulations, the distribution of parameter estimates of \( \sigma(b) \) for such branches shows a large point mass on 0 and the remainder of the probability occurring in a long tail of value > 0. For longer genes, such as RpoC1C2, \( E(\sigma(b)) \) increases, and we are able to obtain meaningful estimates for most branches, with only a few of the very short internal branches being zero, such as those leading to the red algae and the cyanobacteria. For the shortest gene, TufA, many internal branches have a value of \( E(\sigma(b)) \) less than one, which results in little resolution of internal branches in the switching tree because there is little information from which to infer \( \sigma(b) \).

We also examined by eye many of the simulated switching trees and compared them with those estimated from our real data. For the two genes, where we find significant branch-specific switching, AtpA and RpoC1C2, we were confident that the patterns of switching observed in the real data were different from our covarion drift simulation. For the remaining three genes, it was difficult to draw strong conclusions, although switching trees from AtpB and RpoB appeared more different from the covarion drift trees than TufA.

The fifth and sixth columns in figure 2, headed simulated covarion shift, are produced by using simulating data with a change in \( p_{inv} \) on specific lineages using the lineage_specific_seq_gen software (Shavit Grievink et al. 2008). These simulations show the simplest case of a covarion shift, where a single change in \( p_{inv} \) has occurred on the lineage leading to each grouping, where the only associated rate heterogeneity is that introduced by invariable sites. The tree shapes shown in columns 5 and 6 of figure 2 are those inferred under the WAG + F + bCov(I, J), with branch lengths taken to be the median value from 100 simulations. The substitution trees have shapes comparable to those obtained from the real-gene sequences and those from covarion drift simulations.

For AtpA, AtpB, RpoB, and RpoC1C2, the branch lengths in the switching trees are extreme (column 6 of fig. 2), with large amounts of switching occurring on the nonphotosynthetic bacterial lineage and/or the green algal lineage, with the remaining branches tending to have zero length. This observation matches our expectation of what occurs in the simulation, although for RpoB, the lack of a high-switching rate in the bacterial lineage may be surprising. Overall, the switching patterns inferred do not match those from the real data (column 2 of fig. 2), which suggests that a simple switch in the proportion of invariable sites at the base of those groups does not fully describe the heterotachy occurring in the real data. The pattern observed in TufA is different from the other four genes and suggests that there is not enough information in the sequence to infer switching reliably.

**Discussion**

Our novel THMM approach provides a general and intuitive means for investigating rate switching during evolution. The methodology reveals broadly similar conclusions to those proposed in the study of Lockhart et al. (2006), with accelerated rates in the nonphotosynthetic bacteria and the green algal groups being the result of rate switches in those lineages. The precise biological cause of these changes in switching is unclear, but previous research has suggested that changes in the structural and functional constraints acting on proteins could cause such phenomena (Simon et al. 1996; Inagaki et al. 2004). Our new methodology, however, is able to reach these conclusions using a single framework that allows standard hypothesis testing and visualization of where switches have taken place.

**Evidence for Covarion Drift**

Our results show that 4/5 genes examined have significant evidence for some form of heterotachy, either through covarion drift, where rates change gradually and consistently through time, or through covarion shift where there is excessive rate switching on particular branches. This observation extends to amino acid sequences the previous suggestions that heterotachy is a widespread phenomenon in nucleotide evolution (Whelan 2008b) and matches observations made in other studies (e.g., Lopez et al. 2002; Philippe et al. 2003). The TufA gene demonstrated no significant heterotachy of any form, although this result may be due to low statistical power because of the relatively short alignment length of only 235 amino acids.

**Covarion Shifts in Nonphotosynthetic Bacterial and Green Algal Lineages**

In addition to the general heterotachy of the covarion drift model, the AtpA, AtpB, and RpoC1C2 genes, all showed significant evidence for group-specific covarion shifts, whereas only AtpA and RpoC1C2 show additional significant branch-specific covarion shifts. The shapes of the switching trees observed in these four genes suggest that
our results may be conservative. The trees show patterns of switching consistent with a covarion shift occurring in longer branches and fall into two groups. The first group, AtpA and AtpB, shows high levels of substitution and switching in the nonphotosynthetic bacterial lineage, whereas the second group, RpoB and RpoC1C2, shows high levels of substitution and switching in both the nonphotosynthetic bacterial and green algal lineages. These patterns are consistent with the conclusions of Lockhart et al. (2006) and Martin et al. (1998) whereby trees estimated from RNA polymerase genes are different from those estimated from other plastid genes.

Our two simulation studies suggest that the observed patterns of switching may be biologically meaningful. The first simulates data from our null model of covarion drift, where the pattern of switching is expected to be proportional to the amount of substitution on a branch. For AtpA and RpoC1C2, where the branch-specific covarion shift model provided a significant improvement over the covarion drift model, our new method successfully recovered the switching rate in most branches. In all genes, some of the shorter internal branches in the simulation had fewer than a single expected switch occurring. The estimates of the branch-specific switching parameter had a high variance with the majority of the mass on zero switches, resulting in a median estimate of zero. For other branches where several switches were inferred to occur, we successfully estimated the switching rate. These observations suggest that our method tends to be conservative when the number of switches on a branch is small, and we suggest using a combination of a significance test and the pattern before drawing strong conclusions about an individual gene. However, it does not appear that our method is prone to simple artifacts whereby long branches tend to have disproportionately high estimated levels of switching and vice versa.

The second simulation study examines how changes in the proportion of invariable sites in sublineages affect the patterns of switching inferred under our model. These simulations are performed under a simplistic model of change in the proportion of invariable sites that does not include many of the factors occurring in real data and are intended to investigate whether our methodology can recover such changes in simple conditions. We find that our method successfully recovers where such changes occur and for most of the genes the switching trees are similar in their structure to the real data, with AtpA and AtpB having an accelerated switching rate in the nonphotosynthetic bacteria, RpoC1C2 having an accelerated rate in the nonphotosynthetic bacteria and the green algae, and TufA having an unclear pattern of switching attributable to its low information content. The exception to this pattern is RpoB where the real data suggest that it has a pattern similar to RpoC1C2, but the simulation suggesting only the green algae have an accelerated rate of switching. It is not clear why this has occurred, but it may partially be due to the difficulty in recovering estimates of the proportion of invariable sites from four taxa subtrees, which are then subsequently used for simulation. For all the genes, the pattern of switches observed in our simulations is more extreme than the real data, suggesting that the changes in the evolutionary process occurring in the real data are more complex than a simple change in the proportion of invariable sites and that the subtle difference in branch lengths on the switching tree may have biological importance.

Group- and Branch-Specific Parameter Estimation and Hypothesis Testing

Our branch-specific model of covarion shifts introduces a relatively large number of parameters to the evolutionary model, meaning that considerable evolutionary signal is likely to be required to achieve significance. Our results reflect this expectation, with only the long RpoC1C2 gene obtaining significance using standard $\chi^2$ distributions for hypothesis testing, and AtpA also being significant when Monte Carlo simulation is used to generate the null distribution. Interestingly, examining the distributions of the LRT statistic in these null distributions suggests that the $\chi^2$ distribution is conservative in all cases examined. The observation that the null distribution is closer to the asymptotic $\chi^2$ distribution when there are longer sequences (more data) leads us to believe that the lack of fit between our simulations and the expected distributions may disappear for adequately long and diverse sequences.

One method for circumventing the difficulty of testing for covarion shifts using branch-specific switching parameters is to group branches by biological interest a priori. However, this approach introduces two difficulties. First, one must have a clear expectation of which branches should belong within each group. This task is not trivial, even for a well-characterized set of genes such as our plastid–bacterial genes. We find highly significant results for splitting our data into five groups, but interestingly one does not see the same degree of significance if different switching rates are only allowed on the branch leading to each group; in other words, the four branches leading out from the gray anc section in figure 1. This assignment produces nonsignificant comparisons for all but the RpoC1C2 genes, with AtpA, for example, only achieving an increase in likelihood of 1.79 (other data not shown). The second problem with specifying branches a priori is that one may miss other important covarion shifts. Both these factors suggest that the set of branches chosen in a group-specific switching model can have considerable impact on the results obtained.

A distinct advantage of applying branch-specific models is that it allows clear visualization of what is occurring in the data. One must, however, be careful of using such a visualization to obtain a set of branches to feed into a group-specific model because it will invalidate the hypothesis test. However, the addition of simulations can help the biological interpretation of branch-specific results visualizations and the conclusions reached in one set of data may feed forward to other data sets. For example, one could visualize the genes we have examined and hypothesize all plastid genes have acceleration in the green
algal lineage or in the nonphotosynthetic bacterial and green algal lineage. These evolutionary tests could then be performed on subsequent genomic scale data sets to investigate this hypothesis.

Conclusions
This study presents a novel THMM approach that provides a general and intuitive means for investigating rate switching during evolution. The structure of the model is a natural progression from existing covarion-style methods (e.g., Galtier 2001; Huelsenbeck 2002; Wang et al. 2007), allowing progression from a model describing gradual changes in rate over time (covarion drift) to a model that allows all or some subset of lineages to undergo more changes in rate than other lineages (covarion shift) (Ané et al. 2005).

We apply our new model to the data of Lockhart et al. (2006) and, in broad agreement with those authors, find covarion drift in 4/5 genes examined, covarion shifts in the nonphotosynthetic bacteria, and/or the green algal groups for 3/5 genes, and branch-specific covarion shifts in 2/5 genes. Further investigation of the model using simulation demonstrates that the shape of the switching tree and an appropriate hypothesis test can be combined to recover a range of biological scenarios, including accurate recovery of inferred switches when simulating under a null drift model and visual inference of covarion shifts when these are simulated. We also show that the asymptotic $\chi^2$ distributions used for testing hypotheses about models with different parameters on each branch tend to be conservative, which may be the result of having relatively few sites.

The hierarchy of models described here provides a useful tool for investigating changes in evolutionary rates in hitherto unavailable degree of detail. We hope these models and future developments based upon them will be useful for examining rate shifts that result from many important evolutionary events.

Supplementary Material

Supplementary materials are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was started at the Isaac Newton Institute for Mathematical Sciences, Cambridge, in 2007 during the Phylogenetics programme and was further supported by a CoSyst grant from the Linnean Society. B.P.B. is funded by Biotechnology and Biological Sciences Research Council (UK) grant BB/H000445/1. We are grateful to Elizabeth Allman, Felix Feyertag, Chris Howe, and John Rhodes for discussions and two anonymous reviewers for constructive criticism.

References


Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27: 401–410.


