Spatial and Temporal Heterogeneity in Nucleotide Sequence Evolution

Simon Whelan
Faculty of Life Sciences, University of Manchester, Michael Smith Building, Manchester M13 9PT, United Kingdom

Models of nucleotide substitution make many simplifying assumptions about the evolutionary process, including that the same process acts on all sites in an alignment and on all branches on the phylogenetic tree. Many studies have shown that in reality the substitution process is heterogeneous and that this variability can introduce systematic errors into models of phylogenetic analyses. I propose a new rigorous approach for describing heterogeneity called a temporal hidden Markov model (THMM), which can distinguish between among-site (spatial) heterogeneity and among-lineage (temporal) heterogeneity. Several versions of the THMM are applied to 16 sets of aligned sequences to quantitatively assess the different forms of heterogeneity acting within them. The most general THMM provides the best fit in all the data sets examined, providing strong evidence of pervasive heterogeneity during evolution. Investigating individual forms of heterogeneity provides further insights. In agreement with previous studies, spatial rate heterogeneity (rates across sites [RAS]) is inferred to be the single most prevalent form of heterogeneity. Interestingly, RAS appears so dominant that failure to independently include it in the THMM masks other forms of heterogeneity, particularly temporal heterogeneity. Incorporating RAS into the THMM reveals substantial temporal and spatial heterogeneity in nucleotide composition and bias toward transition substitution in all alignments examined, although the relative importance of different forms of heterogeneity varies between data sets. Furthermore, the improvements in model fit observed by adding complexity to the model suggest that the THMMs used in this study do not capture all the evolutionary heterogeneity occurring in the data. These observations all indicate that current tests may consistently underestimate the degree of temporal heterogeneity occurring in data. Finally, there is a weak link between the amount of heterogeneity detected and the level of divergence between the sequences, suggesting that variability in the evolutionary process will be a particular problem for deep phylogeny.

Introduction

Statistical models of nucleotide substitution are widely used to draw evolutionary inferences from alignments of homologous sequence data. Models are used to reconstruct phylogenetic trees (Whelan and Goldman 2001; Felsenstein 2004) and to identify interesting aspects of a sequence’s history, such as changes in selective constraints (Yang and Bielawski 2000) and ancestral sequence reconstruction (Chang et al. 2002). Many commonly used models, such as the family of models nested within the general time reversible model (GTR) (Yang 1994; Whelan et al. 2001), assume that the substitution process is constant between sites in the alignment and throughout the branches of the tree (Kosiol et al. 2006). There is substantial evidence that these assumptions oversimplify the evolutionary process and that not accounting for heterogeneity in sequence evolution can lead to seriously misleading inferences. Consequently, substitution models have been proposed for investigating and describing specific forms of heterogeneity, and their applications to real and simulated data have offered real insights. For example, studies using real data have demonstrated that models that incorporate rate variation across sites provide better tree estimates (Yang 1996; Whelan et al. 2001) and that simple models that fail to account for variation in nucleotide frequencies between sequences in a tree can lead to unreliable phylogenetic tree estimates in Microsporidia (Hasegawa and Hashimoto 1993) and primary endosymbionts (Herbeck et al. 2005). Recently, there has been substantial interest in heterotachy, which occurs when sitewise substitution rates vary between lineages (Lopez et al. 2002; Gruenheit et al. 2008). Simulation studies have shown that data generated by mixtures of the same tree topology with different branch lengths can lead to statistical inconsistency under standard likelihood methods (Kolaczkowski and Thornton 2004). More theoretical studies have shown that such mixtures of branch lengths can result in loss of tree identifiability so that 2 or more trees can equally well describe the observed sequence data (Matsen and Steel 2007). There are currently no general models to account for heterogeneity despite the clear and abundant problems it causes to phylogenetic inference.

For the remainder of this paper, I subdivide heterogeneity into 2 distinct forms:

1. Spatial heterogeneity occurs when characteristics of the substitution process, such as rate or nucleotide frequencies, vary between sites in the sequence. There are many causes of spatial heterogeneity, including the slow/fast pattern of substitution in individual nucleotides in codons.

2. Temporal heterogeneity occurs when characteristics of the substitution process vary in time along the branches of an evolutionary tree. The causes of temporal heterogeneity are usually associated with changes in biological function or selective pressures acting upon a sequence or overall changes in the characteristics of the genome context in which that sequence occurs, such as GC content changes. Heterotachy, for example, is temporal heterogeneity in rate because the rate of evolution changes over time.

Several models have been developed to account for spatial heterogeneity (e.g., Yang 1993; Pagel and Meade 2004) and are typically “mixture models,” which do not assign specific patterns of substitution to specific sites but instead each site has a probability of evolving under one of a selection of models. The most popular mixture model is the use of a $F$ distribution to describe spatial heterogeneity in rate, and most standard models are improved by its inclusion. Other aspects of spatially heterogeneous
evolution, such as variation in nucleotide frequencies or substitution rates between specific nucleotides, have not been so widely investigated. Mixture models have also been developed to account for these forms of spatial heterogeneity in protein evolution (e.g., Lartillot and Philippe 2004), but these methods are not widely used, despite consistently demonstrating that spatial heterogeneity exists. It is possible to further generalize descriptions of spatial heterogeneity using hidden Markov models (HMMs), which have found successful application to protein structure prediction (Thorne et al. 1996) and the identification of conserved elements in aligned genome sequences (Siepel et al. 2005).

There are 2 approaches for describing temporal heterogeneity. The first is to use special cases of the general Markov model (GMM: Barry and Hartigan 1987), which allows for different substitution processes on each branch of the phylogeny. For example, the model of Yang and Roberts (1995) is a special case of GMM that allows the nucleotide composition to vary throughout the tree, whereas the model of Galtier and Gouy (1998) makes further simplifying assumptions and only allows GC content to vary throughout the tree. Derivatives of the GMM have been demonstrated to improve inference on real data (e.g., Herbeck et al. 2005), but their structure makes them impractical for many problems. The number of parameters in the GMM typically grows linearly with the number of branches in the tree, which can make its application to large data sets troublesome. Furthermore, applying such models to tree search can be difficult because each rearrangement in the tree changes the parameters used in the model. An alternative approach to describe temporal heterogeneity is the covarion model (Tuffley and Steel 1996), which allows “switching” between 2 substitution processes through time: 1 where a nucleotide can change (the “on” class) and 1 where the nucleotide cannot (the “off” class). The rate of switching between these classes is defined by the parameters of the model. Several authors have generalized the covarion model by allowing greater variability in the rate of substitution processes and by combining it with mixture models of rate variation. Wang et al. (2007) recently described a generalized model that encompasses the previously published versions of the covarion model as special cases.

There has also been substantial interest in examining heterogeneity in protein evolution to make inferences about structure and function. Sophisticated models have been developed that describe variation in amino acid substitution patterns resulting from changes in residue context (e.g., Soyer and Goldstein 2004) and complex dependencies induced by protein structure (e.g., Rodrigue et al. 2006). Codon substitution models based on the covarion model have also been developed to investigate variation in selective pressures on proteins (e.g., Guindon et al. 2004). These models are usually designed to investigate a single aspect of protein evolution and are therefore difficult to compare to the general models detailed below, although the effects they describe will influence the patterns of nucleotide substitution occurring in protein-coding DNA sequences.

In this study, I quantitatively investigate spatial and temporal heterogeneity in nucleotide composition, transition/transversion substitution bias ($\kappa$), and the overall rate of evolution. I begin by defining a new general form of Markov model, called a temporal hidden Markov model (THMM), which describes heterogeneity as switches between a variety of “hidden” classes, each representing a different type of evolutionary process. A THMM may be considered as a generalization of the covarion model that allows all aspects of the evolutionary process to vary through time, rather than just rate. I proceed to describe how mixture models are special cases of THMMs and that comparisons between different THMMs can be used to investigate the relative importance of different types of spatial and temporal heterogeneity. Application of the models to real coding data reveals that, as expected, the most important factor to include is spatial rate heterogeneity, which is more commonly referred to as rates across sites (RAS). Moreover, failure to explicitly include RAS in the model masks other forms of heterogeneity, with other parameters picking up the slack and attempting to describe RAS as best they can. Once RAS is separately modeled, the relative importance of other forms of heterogeneity is revealed. For these models, accounting for other types of spatial heterogeneity often provides the greatest improvements in fit of the model to the data, with variation in nucleotide composition and $\kappa$ providing a substantially better improvement in fit to data than when further rate variation is included. There is also strong support for models that allow for temporal heterogeneity, with variation in $\kappa$, nucleotide composition, and rate proving progressively more important. I also find that simple approaches for including temporal heterogeneity tend to be equally revealing as more sophisticated methods that detail exactly how different substitution processes change during time. Finally, the most general model examined provides the best description of the evolutionary process for each of the 16 sequence alignments examined.

**Methods**

**A THMM of Evolution**

The family of THMMs described in this study can be considered generalizations of the original covarion model proposed by Tuffley and Steel (1998). Other authors (Galtier 2001; Huelsenbeck 2002) have described more sophisticated versions of the covarion model but all concentrated on solely modeling rate heterogeneity. Recently, Wang et al. (2007) have combined all 3 versions into a general covarion model (GCM) and applied it to amino acid sequences, where the original interpretation of the covarion model is more fitting. In contrast, the aim of this study is to introduce and apply a new form of model to understand the relative importance of spatial and temporal heterogeneity at the nucleotide level.

For consistency, I shall, where possible, follow the notation of the GCM of Wang et al. (2007) and, where appropriate, highlight differences between my proposed general THMMs and the GCM. A THMM consists of $g$ separate substitution processes, with the $l$th process represented by a $4 \times 4$ instantaneous rate matrix, $M_l^{\kappa}$, with elements $M_{ij}^{\kappa}$. Each constituent subprocess is the equivalent of a standard nucleotide substitution model and makes the assumptions of reversibility, stationarity, and homogeneity. In
this study, for simplicity, each $M_k$ is a separate HKY model so that
\[
\mathbf{M}^k = \mu^k \left[
\begin{array}{cccc}
-\tilde{\pi}_l^k & \kappa^k \tilde{\pi}_G^k & \tilde{\pi}_T^k \\
\pi_A^k & -\tilde{\pi}_G^k & \pi_T^k \\
\kappa^k \tilde{\pi}_A^k & \kappa^k \tilde{\pi}_C^k & -\tilde{\pi}_G^k \\
\pi_A^k & \pi_C^k & \pi_G^k
\end{array}
\right].
\]
(1)

For the $k$th process, the parameters $\pi_A^k, \pi_C^k, \pi_G^k,$ and $\pi_T^k$ define the nucleotide frequencies, the parameter $\kappa^k$ describes the transition/transversion rate ratio, and the parameter $\mu^k$ describes the overall substitution rate of the process. This parameterization contrasts with the GCM model, which allows only $\mu$ to vary between the substitution processes. In keeping with the HMM nomenclature, each subprocess is kept hidden class in a standard HMM.

The $g$ classes are linked by a reversible switching process that describes the rate of switching between the different sub-processes, described by the $g \times g$ instantaneous rate matrix:
\[
\mathbf{C} = \left[
\begin{array}{cccc}
-\tilde{\pi}_1 & \rho^{1.2} \tilde{\pi}_2 & \ldots & \rho^{1.6} \tilde{\pi}_6 \\
\rho^{1.2} \tilde{\pi}_1 & 0 & \ldots & \rho^{2.6} \tilde{\pi}_6 \\
\vdots & \vdots & \ddots & \vdots \\
\rho^{1.6} \tilde{\pi}_1 & \rho^{2.6} \tilde{\pi}_2 & \ldots & -
\end{array}
\right].
\]
(2)

The parameters $\tilde{\pi}_1, \tilde{\pi}_2, \ldots, \tilde{\pi}_6$ define the probabilities of each hidden class, and the parameter $\rho^{k,l}$ defines the relative rate of switching between hidden classes $k$ and $l$. When the whole set of $\rho^{k,l}$ are estimated from the data they will, hereafter, be referred to as $\rho$, in reference to their similarity to the exchangeabilities in the GTR nucleotide model. The $C_{k,l}$ entries are equivalent to the rates of transition between hidden classes in a standard HMM.

These $\mathbf{M}$ and $\mathbf{C}$ matrices can be combined to form the $4g \times 4g$ instantaneous rate matrix, $\mathbf{Q}$, for the THMM with the off-diagonal elements defined by
\[
Q_{i,j} = \begin{cases} 
M_{i,j} & \text{for all } i \neq j \text{ and } k = l \\
\pi_k \rho^{k,j} & \text{for all } i = j \text{ and } k \neq l \\
0 & \text{for all } i \neq j \text{ and } k \neq l
\end{cases}
\]
(3)

Nucleotide substitutions are described by changes between states $i$ and $j$, whereas switches between hidden classes are described by changes between $k$ and $l$. In this notation, hidden classes are referred to by superscripts, and observable states (nucleotides) are referred to by subscripts. Note that the inclusion of the term $\pi_k$ with $C_{k,l}$ is required to ensure that the model satisfies the detailed balance equations, whereby $\pi_k Q_{i,j} = \pi_l Q_{j,i}$, therefore defining the model as reversible. Definitions of covarion-like models to date, including the GCM, do not include this term and therefore could not allow nucleotide content to change through time. The stationary distribution for the $i$th nucleotide of process $k$ occurring is $\pi_i^k = \tilde{\pi}_i \pi_k$. The diagonals are defined as the negative row sums in the usual manner, ensuring that the instantaneous rate matrix generates a valid family of probability matrices.

In order for the branch lengths to be meaningful, all entries in $\mathbf{Q}$ describing substitutions between observable states (nucleotides) are scaled so that the mean rate of switching in the process is 1 so that $\sum_i \sum_j \pi_j Q_{i,j} = 1$. This scaling is in common with all other substitution models and effectively removes a single degree of freedom from the model. Another quantity of interest to calculate is the rate of switching between hidden classes relative to the rate of substitution between nucleotides. Given the $\mathbf{Q}$ matrix is scaled to 1, the relative rate of switching can be calculated as $T_{rate} = -\sum_i \sum_j \pi_j Q_{i,j}$.

For some models, the THMM will be combined with the standard $\Gamma$-distributed RAS approach with 4 discrete rate categories, which independently describe spatial rate variation. This approach incorporates into the model a single additional parameter, $\alpha$, the value of which is inversely related to the amount of spatial rate heterogeneity in the sequences. When applying $\Gamma$-distributed RAS, the whole of $\mathbf{Q}$ is scaled by the appropriate rate, resulting in both the substitution rate and $T_{rate}$ varying between rate categories.

Modeling Spatial and Temporal Heterogeneity

Modeling nucleotide substitution using THMMs incorporates both spatial and temporal heterogeneity. Teasing apart the relative contributions of these 2 forms of heterogeneity offers insights into both how the evolutionary process shapes sequences and the kind of components we need to include in substitution models to adequately describe this heterogeneity. THMMs are ideally suited to this purpose and placing a variety of restrictions on them offers different insights into different forms of heterogeneity. The first interesting restriction is to set all switching parameters, $\rho^{k,l}$, to zero, completely removing the ability of THMMs to describe temporal heterogeneity. This restriction reduces the THMM to the standard form of phylogenetic mixture model used to describe spatial heterogeneity, where the probability of each individual hidden process occurring at a site is defined by the probabilities $\tilde{\pi}_l$. The importance of spatial heterogeneity is therefore measured as the increase in likelihood obtained by the mixture model over the standard HKY model. Restricted forms of mixture model, where only one type of parameter is allowed to vary between classes, are investigated to examine the relative importance of spatial heterogeneity in rate, nucleotide composition, and transition/transversion substitution bias. The sum increase in likelihoods of these restricted models is also compared with that of the full mixture model to investigate whether the restricted models are describing the same phenomenon.

The second interesting restriction to the THMM is to force the switching parameters of $\rho^{k,l}$ to take an equal value but allow this single parameter, $\rho$, to be estimated from the data. Comparing the increase in likelihood of $+\rho$ THMMs relative to mixture models provides a measure of the importance of incorporating temporal heterogeneity into the substitution process. Restricted forms of the model are also studied to assess the relative importance of temporal heterogeneity in rates, nucleotide composition, and transition/transversion substitution bias. All these restricted $+\rho$ THMMs are compared with a full THMM where $\rho^{GTR}$ is free to vary and estimated from the data. Comparisons between $+\rho$ and $+\rho^{GTR}$ models measure
the importance of allowing different rates of switching between hidden classes.

It is interesting to note that under THMMs, it is impossible to investigate temporal heterogeneity without also incorporating spatial heterogeneity because the THMM structure assumes no knowledge of when and where switches in class occur, with each site and position in the tree being equally likely. Therefore, a THMM can approximate a nonstationary, nonreversible process by assuming that at any point in time, the same proportion of sites evolve to each of the hidden classes but exactly which sites evolve to which classes are unknown and varies through time. This approach could be highly effective at describing broad heterogeneous effects, such as heterotachy, but may not be as effective as the GMM for cases where the sites are highly correlated at each point in time, such as when GC content gradually changes through a tree.

Technical Details, Computation, and Optimization

Likelihood computations are performed in the usual manner (Felsenstein 1981), and the term likelihood in the text is used to refer to the log likelihood of the model. All model parameters are estimated from the data using standard numerical optimization techniques. Parameter estimation can be troublesome under some models and multiple local optima were identified, particularly in models that have varying nucleotide frequencies between different hidden classes. Further investigation has confirmed that this is only a numerical optimization problem: the model is formulated so that no 2 hidden processes can be the same and more formal work has confirmed that 3-class THMMs are identifiable (Allman E, Rhodes J, unpublished data). Standard techniques were used to address the multiple optima problem, including starting all optimizations from multiple starting points and “shuffling” the combinations of labels of different parameters. All calculations were performed using Leaphy (an acronym for likelihood estimation algorithms in phylogenetics; available at http://www.manchester.ac.uk/bioinformatics/leaphy), a software program developed and distributed by the author (Whelan 2007). All models discussed in this paper will be incorporated into a future release of Leaphy.

Model comparison is performed using an Akaike (1974) information criterion (AIC), which is calculated for model $M$ as $\text{AIC}(M) = 2k - 2\ln\hat{L}(M)$, where $k$ is the number of parameters in the model and $\ln\hat{L}(M)$ is the maximum log likelihood of model $M$. The results presented compare AIC between the model of interest, which incorporates some form of heterogeneity, and an appropriate baseline model. The increases in AIC relative to the baseline model are compared for different models and are interpreted as relative improvement in fit of the model to the observed sequence data, with higher scores indicating better fit.

Sequence Data and Trees

In this study, 16 different sequence alignments are used to investigate spatial and temporal heterogeneity in nucleotide substitution. The first part of the study investigates heterogeneity in the groEL data of Herbeck et al. (2005) where, in contrast to the original study, all 3 codon positions are included in the analysis. The groEL gene is also of interest because previous studies have suggested that there is substantial coevolution acting between the residues in the protein, which may manifest itself as additional evolutionary heterogeneity (Fares and Travers 2006). The second part of the study generalizes these observations by investigating an addition 15 sequence alignments from the Pandit database (Whelan et al. 2006; available at http://www.ebi.ac.uk/goldman-srv/pandit/), chosen to have long alignments with a range of divergences and number of sequences. A summary of these data is provided in table 1, demonstrating a wide range of protein domains, with different sequence characteristics, involved in a variety of biochemical functions.

### Table 1

Summary of Sequence Alignments Used in This Study

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Name</th>
<th>Sequences</th>
<th>Length</th>
<th>TL(HKY)</th>
<th>TL(HKY + 1)</th>
<th>%GC</th>
<th>Standard Deviation</th>
<th>dN/dS</th>
</tr>
</thead>
<tbody>
<tr>
<td>groEL</td>
<td>Heat shock 60-kDa protein 1</td>
<td>23</td>
<td>1,572</td>
<td>2.76</td>
<td>8.08</td>
<td>0.49</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>PF00294</td>
<td>pKB family carbohydrate kinase</td>
<td>77</td>
<td>972</td>
<td>42.40</td>
<td>56.14</td>
<td>0.52</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>PF00450</td>
<td>Serine carboxypeptidase</td>
<td>81</td>
<td>1,497</td>
<td>25.9</td>
<td>34.77</td>
<td>0.46</td>
<td>0.07</td>
<td>0.11</td>
</tr>
<tr>
<td>PF00503</td>
<td>G-protein $\alpha$ subunit</td>
<td>24</td>
<td>1,083</td>
<td>9.46</td>
<td>13.17</td>
<td>0.45</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>PF00759</td>
<td>Glycosyl hydrolase family 9</td>
<td>38</td>
<td>1,530</td>
<td>12.71</td>
<td>19.60</td>
<td>0.49</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>PF00873</td>
<td>AcrB/AcrD/AcrF family</td>
<td>23</td>
<td>3,198</td>
<td>10.43</td>
<td>15.73</td>
<td>0.53</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td>PF00888</td>
<td>Cullin family</td>
<td>48</td>
<td>1,968</td>
<td>30.93</td>
<td>46.68</td>
<td>0.42</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>PF00999</td>
<td>Sodium/hydrogen exchanger family</td>
<td>83</td>
<td>1,353</td>
<td>40.57</td>
<td>56.05</td>
<td>0.49</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>PF01212</td>
<td>Beta-eliminating lyase</td>
<td>52</td>
<td>1,191</td>
<td>20.96</td>
<td>21.76</td>
<td>0.54</td>
<td>0.14</td>
<td>0.08</td>
</tr>
<tr>
<td>PF01501</td>
<td>Glycosyl transferase family 8</td>
<td>70</td>
<td>1,026</td>
<td>27.47</td>
<td>29.52</td>
<td>0.42</td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>PF01602</td>
<td>Adaptin N terminal region</td>
<td>33</td>
<td>1,890</td>
<td>16.80</td>
<td>25.30</td>
<td>0.45</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>PF01757</td>
<td>AcrB/AcrD/AcrF family</td>
<td>60</td>
<td>1,377</td>
<td>29.19</td>
<td>46.91</td>
<td>0.47</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>PF01979</td>
<td>Amidohydrolase family</td>
<td>64</td>
<td>1,437</td>
<td>31.96</td>
<td>44.35</td>
<td>0.50</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>PF07394</td>
<td>Protein of unknown function</td>
<td>51</td>
<td>1,368</td>
<td>30.56</td>
<td>46.68</td>
<td>0.47</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>PF07596</td>
<td>Protein of unknown function</td>
<td>74</td>
<td>1,077</td>
<td>19.96</td>
<td>27.92</td>
<td>0.57</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>PF07748</td>
<td>Glycosyl hydrolases family 38 C-terminal domain</td>
<td>46</td>
<td>1,902</td>
<td>22.10</td>
<td>26.72</td>
<td>0.47</td>
<td>0.11</td>
<td>0.06</td>
</tr>
</tbody>
</table>

*TL refers to tree length, which is the sum of branches in the phylogenetic tree.

$^a$ $dN/dS$ calculated using codeml (Yang 1997) with $F3 \times 4$ frequencies.
Results and Discussion

Heterogeneity in groEL Data without Γ-Distributed RAS

The increase in (log) likelihood and AIC of the new mixture models and THMMs relative to the simple HKY model are shown in table 2. The rows describe models with different quantities varying between classes, whereas the columns describe different forms of model and number of hidden classes. The ΔAIC values show that the best fit to the data was achieved by the 3-class version of the most general model, All(3) + ρ\textsubscript{GTR}, which allows all forms of spatial and temporal heterogeneity. The following discusses how specific forms of heterogeneity describe different parts of the evolutionary process.

Spatial Heterogeneity

The 2 columns under mixture models in table 2 describe the improvement in fit observed by adding spatial heterogeneity to the HKY model. The difference in values between the 2 columns represents the improvement in fit of the 3-class mixture model relative to the 2-class model. In all cases, the 3-class model provides a better fit than the 2-class model, although for the Kappa models the increase in likelihood (ΔAIC) of 316.6 − 299.1 = 17.5 (533.2 − 500.3 = 32.9) is relatively modest. The models that allow only individual types of spatial heterogeneity indicate that the rate variation is, as expected, the single most important form of heterogeneity (Rates model), followed by nucleotide frequencies (Frequencies model), and then variation in transition/transversion substitution bias (Kappa model). In agreement with other studies, estimating the rates and relative proportions of different hidden classes in the Rates(3) mixture model (ΔAIC = 4784.3) provides a small, but significant, better fit to the data than using Γ-distributed RAS (ΔAIC = 4643.4; calculated from the 2369.7 improvement in likelihood obtained by adding Γ-distributed RAS to HKY). The model containing all types of spatial variation provides a better fit than the models incorporating single types of heterogeneity, suggesting that there is substantial spatial heterogeneity that cannot be described with a simple model where a single parameter varies between sites.

Temporal Heterogeneity

The 2 columns under “single switching parameter (ρ)” in table 2 describe the improvement in fit observed by incorporating temporal heterogeneity to the model by adding a single switching parameter to describe all switches between classes. These values, when compared with their corresponding value under the mixture model, show limited support for temporal heterogeneity. In some cases, the model without temporal heterogeneity is preferable, for example, the All(2) + ρTHMM confers no improvement in likelihood over the All(2) mixture model (2353.3 for both models), and, consequently, the simpler mixture model has a better ΔAIC score. In other cases, usually for the 3-class models, incorporating temporal heterogeneity gives a small increase in likelihood (AIC), with the largest increase in likelihood (AIC) of 2694.0 − 2673.2 = 20.8 (5266.0 − 5226.3 = 39.7) observed by incorporating ρ\textsubscript{GTR} into the All(3) model. The results in table 2 show that spatial rate variation explains a large proportion of the increase in likelihood obtained for the All models: the Rates(3) model, for example, provides 91% (4784.3/5266.0) of the ΔAIC increase in fit observed by the most general All(3) + ρ\textsubscript{GTR} model. Interestingly, these results contrast with those obtained by Herbeck et al. (2005), who found significant GC variation in the data, enough to affect the inferred tree topology and bias their results. This suggests that the strong signal for spatial rate variation masks other effects from the THMM.

Model Details

Examining the instantaneous rate matrices (\textbf{Q}) for these models further demonstrates the supreme importance of spatial heterogeneity in rate. The top left plot of figure 1 shows \textbf{Q} for Frequencies(3) + ρ has a most peculiar form: a nucleotide evolving under this model would spend approximately 65% of its time as a G nucleotide in the third hidden class. This unusual pattern can be interpreted as the model attempting to describe spatial rate variation through the adjustment of nucleotide frequencies. The high probability of visiting and remaining in G is an approximation to an invariant site model, where each column in an alignment site has a probability of being an invariant G. The \textbf{Q} matrix for All + ρ\textsubscript{GTR}, shown in the bottom left, is more reasonable, although it tends to describe spatial and temporal rate variation, with the classes describing progressively faster evolving sites. There is also evidence of temporal variation in rate (heterotachy), with a relatively high rate of switching between the first and third class.

Heterogeneity in groEL Data Including Γ-Distributed RAS

The results presented above agree with other studies showing that spatial heterogeneity in rate is of great
importance in sequence evolution (Yang 1996), which warrants its separate inclusions in the model. I account for spatial rate variation using a $\Gamma$ distribution, with each site having an equal probability of evolving to 1 of 4 THMM processes, each with different rates. Table 3 shows the increase in likelihood ($\Delta$AIC) relative to the HKY + $\Gamma$ model for incorporating different types of spatial and temporal heterogeneity. In common to modeling heterogeneity without RAS (above), the most complex model, All(3) + $\Gamma$, provides the best description of the data. In contrast to models without RAS, all ways of adding complexity provide a substantially improved model fit, although the patterns of improvement are complex.

Spatial Heterogeneity

The largest improvements in model fit are observed when incorporating spatial heterogeneity. The single most important factor is spatial heterogeneity in $\kappa$, with Kappa(2) + $\Gamma$ providing the single largest increase in likelihood ($\Delta$AIC) of 261.0 (518.0). Adding an additional class to Kappa(2) + $\Gamma$ provides a much smaller improvement in likelihood ($\Delta$AIC) 286.6/C0 261.0 5 25.6 (567.2/C0 518.0 5 49.2). The next most important factor is spatial heterogeneity in nucleotide frequencies, with Frequencies(2) + $\Gamma$ capturing most of the improvement in likelihood. In contrast to Kappa, however, progressing to a 3-class model also provides a large improvement in likelihood ($\Delta$AIC) of 286.6 − 261.0 = 25.6 (567.2 − 518.0 = 49.2). Incorporating spatial rate heterogeneity over and above $\Gamma$-distributed RAS provides some improvement in fit but less so than other forms of spatial heterogeneity. As before, combining all forms of spatial heterogeneity into the All(3) + $\Gamma$ model provides the greatest improvement in likelihood ($\Delta$AIC) improvement of HKY + $\Gamma$.

Fig. 1.—Bubble plots illustrating the instantaneous rate matrix $Q$ for different THMMs, with bubble area proportional to the appropriate element in the matrix. Numbers to the top right and top left of the plots describe the rate of switching between hidden classes relative to the rate of substitution.
Comparing the sum of likelihoods of the individual types of heterogeneity (Rates, Frequencies, and Kappa models) with the improvement observed when combining them (All models) reveals how the different components of the models interact. For the 2-class mixture model, the improvement in likelihood of 278.6 of All(2) + Γ over HKY + Γ is around half (53%) the sum of the improvement of the separate model components (69.5 + 196.6 + 261.0 = 527.1), indicating either that parameters are describing the same phenomenon or that a single effect dominates. The All(3) + Γ model provides an improvement in likelihood over HKY + Γ of 433.3, which is a larger proportion (66%) of the sum of its components (79.1 + 282.7 + 286.6 = 648.4), suggesting that including more hidden classes allows the model to account for additional types of spatial heterogeneity.

**Temporal Heterogeneity**

There is evidence for incorporating temporal heterogeneity with a single ρ parameter for both the 2 and 3 hidden class models, with improvements in likelihood (AIC) ranging from 442.8 – 433.3 = 9.5 (859.6 – 842.5 = 17.1) for All(3) + ρ + Γ to 338.1 – 282.7 = 55.4 (658.1 – 549.3 = 108.8) for Frequencies(3) + ρ + Γ, with an average across all models of 27.9 (53.8). All these increases are highly significant in a hypothesis testing framework (likelihood ratio tests; P < 0.001): taking the mixture model as the null hypothesis and the THMM as the alternative model, the 2 hypotheses differ by a single degree of freedom, and a $\chi^2$ distribution provides a conservative test of whether temporal heterogeneity occurs in the data (Self and Liang 1987; Whelan and Goldman 1999). Allowing more complex switching patterns between hidden classes with ρ parameters also provides a large improvement in likelihood and AIC over having a single ρ parameter. The observations of temporal heterogeneity in nucleotide composition now match those of Herbeck et al. (2005) but demonstrate that additional heterogeneity is also present in their data.

**Model Details**

The Q matrices for the Frequencies(3) + ρ + Γ and All(3) + ρGTR + Γ model are shown in the right-hand portion of figure 1. In contrast to Frequencies(3) + ρ, the patterns of substitution estimated from Frequencies(3) + ρ + Γ appear reasonable and have no peculiar dominant effects. For All(3) + ρGTR + Γ, there is much more variation between the rates matrices of the hidden classes, relative to the All(3) + ρGTR model. The shape and form of these matrices are particularly interesting and can be interpreted as reflecting different aspects of the genetic code. The first hidden class ($\approx 0.32$) represents slowly evolving nucleotides (scaled rate = 0.07) with an unexpected bias toward transition substitutions ($\kappa = 0.10$). This class may be interpreted as representing the slowly evolving residues in the protein, which would frequently include cases when both the first and second positions are highly restricted and cases where the third position is in a 1-fold degenerate codon, such as tryptophan or methionine. The second hidden class (0.34) is more quickly evolving (scaled rate = 1.63) and has a bias toward transition substitutions similar to that which occur in noncoding DNA ($\kappa = 1.56$). For All(3) + ρGTR + Γ, the pat-

<table>
<thead>
<tr>
<th>Number of Hidden Classes</th>
<th>Mixture Models</th>
<th>Single Switching Parameter (ρ)</th>
<th>ρGTR Switching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rates + Γ</td>
<td>69.5 (135.1)</td>
<td>84.1 (162.3)</td>
<td>—</td>
</tr>
<tr>
<td>Frequencies + Γ</td>
<td>196.6 (385.1)</td>
<td>234.7 (459.5)</td>
<td>—</td>
</tr>
<tr>
<td>Kappa + Γ</td>
<td>261.0 (518.0)</td>
<td>291.2 (576.3)</td>
<td>—</td>
</tr>
<tr>
<td>All + Γ</td>
<td>278.6 (545.3)</td>
<td>302.8 (591.6)</td>
<td>456.4 (884.9)</td>
</tr>
</tbody>
</table>

* Log likelihood under HKY + Γ is $-16209.9$ and improvement of HKY + Γ over HKY is $2369.7$. The values in bold indicate the best fitting model under AIC.

Patterns of Spatial and Temporal Heterogeneity from Multiple Data

**General Patterns of Heterogeneity**

The overall results of including spatial and temporal heterogeneity in the groEL alignment and data taken from Pandit are shown in tables 4 and 5. The overwhelming...
importance of spatial rate heterogeneity is again demonstrated in table 4; the improvement in likelihood observed in All(3) + ρGTR is only a fraction higher than the simpler Rates(3) + ρ model, which only includes rate heterogeneity. Adding Γ-distributed RAS to the model reveals new insights into the types of heterogeneity present in the data, with All(3) + ρGTR + Γ yielding an increase in likelihood of 5235.9 over All(3) + ρGTR. It therefore seems appropriate to generalize the observation above that spatial rate heterogeneity masks other forms of heterogeneity, unless explicitly accounted for in an evolutionary model. This strongly suggests that even quite sophisticated tests may seriously underestimate the importance of temporal heterogeneity unless RAS is carefully and separately incorporated into the model.

The results from THMMs with Γ-distributed RAS are summarized in table 5 and provide more insights into other forms of heterogeneity. For all 15 Pandit families, the most complex THMM, All(3) + ρGTR + Γ, provides the best description of the data according to AIC (data not shown), with improvements ranging from a minimum of 1041.7 for PF00503 (G-protein subunit family) to a maximum of 35438.4 for 2-class models. Temporal heterogeneity is also an important factor in nucleotide frequencies and, to a lesser extent, kappa.

**Specific Patterns of Heterogeneity in Γ-Distributed RAS Models**

The relative importance of different types of heterogeneity varies from family to family, which is expected because each of the families has evolved under different selective pressures. However, many of the observations made in the groEL data generalize to all families. On average, adding spatial heterogeneity in either Frequencies or Kappa provides substantial improvements in model fit, whereas only a relatively small, although significant, improvement in likelihood is observed by incorporating additional spatial heterogeneity in rate. The relative importance of including spatial and temporal heterogeneity also varies by the number of hidden classes in the model. For 2-class models, the relative importance of spatial heterogeneity in nucleotide frequencies (∆AIC = 8140.7) and kappa (∆AIC = 8307.9) is similar, with 9/16 families having a higher ∆AIC under Frequencies(2) + Γ than Kappa(2) + Γ. For 3-class models, the picture is different: a class added to Kappa(2) + Γ provides a total improvement in likelihood (∆AIC) of 4795.0 – 4186.0 = 609.0 (9491.1 – 8307.9 = 1183.2), whereas an additional class in the Frequencies(2) + Γ model provides a smaller total improvement in likelihood (∆AIC) of 6293.6 – 4134.3 = 2159.3 (12331.2 – 8140.7 = 4190.5).

There is also strong evidence for temporal heterogeneity in all families. On average, the most important form of temporal heterogeneity is that of rate (heterotachy), which provides a large improvement in model fit, with ∆AIC = 10025.3 – 2613.9 = 7411.4 for 2-class models and ∆AIC = 10401.7 – 3360.0 = 7041.7 for 3-class models. Temporal heterogeneity is also an important factor for nucleotide frequencies and, to a lesser extent, kappa. The improvement in likelihood (∆AIC) obtained by adding temporal heterogeneity in frequencies to 2-class and 3-class mixture models was 6594.5 – 4134.3 = 2460.2 (13029.1 – 8140.7 = 4888.4) and 9629.7 – 6293.6 = 3336.1 (18971.3 – 12331.2 = 6640), respectively. The progressive increase in likelihood and model fit suggests that many hidden classes may be required to adequately

### Table 4

**Overall Improvement in Log Likelihood (∆AIC) for 16 Data Sets Relative to HKY* for Models with No RAS**

<table>
<thead>
<tr>
<th>Number of Hidden Classes</th>
<th>Mixture Models</th>
<th>Single Switching Parameter (p)</th>
<th>ρGTR Switching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Rates</td>
<td>30688.7 (61219.4)</td>
<td>36012.5 (71835.0)</td>
<td>35508.0 (70826.1)</td>
</tr>
<tr>
<td>Frequencies</td>
<td>9807.4 (19392.8)</td>
<td>15461.8 (30573.6)</td>
<td>12984.6 (25715.1)</td>
</tr>
<tr>
<td>Kappa</td>
<td>3860.6 (7563.1)</td>
<td>4784.5 (9379.1)</td>
<td>5140.8 (10091.5)</td>
</tr>
<tr>
<td>All</td>
<td>31958.1 (63630.2)</td>
<td>39035.5 (77593.1)</td>
<td>36760.4 (71835.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The total log likelihood under HKY is −1053026.8.

The values in bold indicate the best fitting model under AIC.

### Table 5

**Overall Improvement in THMMs Log Likelihood (∆AIC) for 16 Data Sets Relative to HKY + Γ for Models with Γ-Distributed RAS**

<table>
<thead>
<tr>
<th>Number of Hidden Classes</th>
<th>Mixture Models</th>
<th>Single Switching Parameter (p)</th>
<th>ρGTR Switching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Rates + Γ</td>
<td>1338.9 (2613.9)</td>
<td>1728.0 (3360.0)</td>
<td>5060.7 (10025.3)</td>
</tr>
<tr>
<td>Frequencies + Γ</td>
<td>4134.3 (8140.7)</td>
<td>6293.6 (12331.2)</td>
<td>6594.5 (13029.1)</td>
</tr>
<tr>
<td>Kappa + Γ</td>
<td>4186.0 (8307.9)</td>
<td>4795.0 (9491.1)</td>
<td>5705.7 (11315.5)</td>
</tr>
<tr>
<td>All + Γ</td>
<td>6379.5 (12567.0)</td>
<td>9299.0 (18214.0)</td>
<td>9265.1 (18306.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The total log likelihood under HKY + Γ is −1017588.4 and improvement of HKY + Γ over HKY is 35438.4.

The values in bold indicate the best fitting model under AIC.
explain temporal heterogeneity in nucleotide composition. In contrast, the improvement observed for adding temporal heterogeneity to 2-class and 3-class Kappa models was quite similar at 5705.7 − 4186.0 = 1519.7 (11315.5 − 8307.9 = 3007.6) and 6303.6 − 4795.0 = 1508.6 (12479.2 − 9494.1 = 2985.1), respectively. This observation suggests that temporal variation in \( \kappa \) could be quite limited and, potentially, is adequately described by a small number of hidden classes for many data sets. It is also interesting to note that the improvement in model fit obtained by incorporating spatial heterogeneity in Frequencies and Kappa is often higher than that obtained by incorporating heterotachy, and in 8/16 families, the Frequencies model provided a larger increase in \( \Delta AIC \) than incorporating heterotachy.

Exploratory analysis with more complex THMMs with 4 hidden classes found further large improvements were possible for all data sets, although these data are not shown due to numerical optimization problems.

Model Details

The patterns of evolution observed in \( Q \) under All(3) + \( \rho^{GTR} + \Gamma \) for the different families broadly agree with those observed in the groEL data. There tends to be one class that is dominated by a large value of \( \kappa \) and, interestingly, the models without temporal heterogeneity (mixture models) exhibit this effect but to a much weaker degree. This observation strongly suggests that some sites in codon sequences only accept transition substitutions during part of their evolution, but few sites evolve in this manner for the whole time. It also explains why a single hidden class seems adequate to describe variation in \( \kappa \) through a tree. The remaining 2 classes tend to vary primarily in rate and nucleotide composition, with limited distinguishing features. One class does tend to be slow and the other fast, suggesting they contribute to describing heterotachy, and the classes also tend to vary in their nucleotide composition. However, the lack of clearly interpretable patterns implies that the 2 other classes are not describing a single phenomenon. Instead I hypothesize that these 2 classes are capturing general aspects of the genetic code, in addition to other types of heterogeneity. Informed restructuring of the model by adding carefully parameterized hidden classes may reveal further details about the different forms of heterogeneity in the data.

Table 6 shows the rate of switching between hidden classes relative to the substitution rate (and standard deviation) across the 16 Data Sets Examined.

<table>
<thead>
<tr>
<th>Number of Hidden Classes</th>
<th>No RAS</th>
<th>( \Gamma )-Distributed RAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rates</td>
<td>0.12 (0.05)</td>
<td>0.11 (0.05)</td>
</tr>
<tr>
<td>Frequencies</td>
<td>0.16 (0.20)</td>
<td>0.14 (0.10)</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.17 (0.05)</td>
<td>0.22 (0.08)</td>
</tr>
<tr>
<td>All</td>
<td>0.13 (0.06)</td>
<td>0.15 (0.08)</td>
</tr>
<tr>
<td>All + ( \rho^{GTR} )</td>
<td>—</td>
<td>0.21 (0.13)</td>
</tr>
<tr>
<td>All + ( \rho^{GTR} + \Gamma )</td>
<td>—</td>
<td>0.30 (0.14)</td>
</tr>
</tbody>
</table>

Tree Length Correlates with Heterogeneity

The existence of the genetic code in protein-coding sequences appears to heavily contribute to the observed heterogeneity. Figure 2 shows the increase in \( \Delta AIC \) per site for the models plotted against the tree length under HKY(±\( \Gamma \)) for each family. There is an apparent correlation between these variables, with longer trees tending to demonstrate a greater degree of heterogeneity. For example, the \( \times \) symbol in the bottom right-hand graph shows the increase in model fit per site for the All(2) + \( \Gamma \) mixture model, and the hollow squares (\( \square \)) show the same value for the All(3) + \( \rho + \Gamma \) THMM. Both show linear correlations with tree length (All(2) + \( \Gamma \), \( R^2 = 0.20 \); All(3) + \( \rho + \Gamma \), \( R^2 = 0.15 \)), although closer inspection shows that the \( R^2 \) values induced by the regression may be spurious because there appear to be 2 different types of improvement. There appears to be a relatively steep increase in \( \Delta AIC \) per site for 6 points on the graph between tree lengths in the range of 10–40; for the remaining 10 points, ranging from a tree length of 10–65, there seems to be a more gradual increase in model fit. Plotting the increase in \( \Delta AIC \) observed by progressing from the mixture model to the THMM also visually correlates with tree length, although the division into 2 groups results in a low \( R^2 \) value of 0.05. This separation into 2 groups is apparent in all the graphs shown in figure 2, although it is more pronounced for models with \( \Gamma \)-distributed RAS. I have found no reasonable explanation for this observation, either biologically or model related, and it remains an open question as to what causes the 2 groupings.

Conclusions

I present a general THMM and apply it to investigate spatial and temporal heterogeneity in nucleotide sequence evolution. Each hidden class in the THMM can adapt to heterogeneity in rate, nucleotide frequencies, or bias toward transition mutations, or any combination of the 3. Simplified versions of the model that only incorporate a single type of heterogeneity reveal extensive spatial rate
heterogeneity and, unless this is explicitly incorporated elsewhere, other parameters in the THMM all attempt to describe RAS. In all data sets examined, adding standard $\Gamma$-distributed RAS to the THMM reveals a range of other significant heterogeneity, leading me to conclude that attempts to quantify different forms of heterogeneity, such as those of Ane et al. (2005), may be limited unless they include explicit modeling of spatial rate heterogeneity.

All data sets examined exhibit pervasive spatial and temporal heterogeneity. Very large increases in likelihood and model fit can be obtained by adding relatively few parameters to existing models to describe heterogeneity, which suggests that simple homogeneous models, such as HKY($\Gamma$) and GTR($\Gamma$), do not appear to adequately encapsulate naturally occurring heterogeneity. In turn, this could result in biased estimates of any evolutionary parameters inferred. This is known to be the case for spatial and temporal rate heterogeneity (e.g., Yang 1995; Lopez et al. 2002; Gruenheit et al. 2008), and it seems reasonable to assume that other forms of heterogeneity will induce the same effects. The degree of heterogeneity exhibited by data sets appears weakly correlated with the amount of evolution that has occurred between the sequences (tree length). This correlation, and the large increases in model fit obtained by

![Figure showing the increase in AIC per site of different THMM models relative to an appropriate baseline model. The label above each graph describes the parameters allowed to vary in the more complex models. Two hidden class models ($q = 2$) are described by $\times$ for a mixture model and $+$ for a THMM. Three hidden class models ($q = 3$) are described by $\Diamond$ for a mixture model and $\square$ for a THMM. Note the change in scale on the $x$ and $y$ axis between models with and without $\Gamma$-distributed rate variation.]
incorporating heterogeneity into substitution models, strongly suggests that more sophisticated models of evolution are required, particularly if we wish to accurately infer deep phylogenetic relationships from distantly related sequences.

Interestingly, it seems that a proportion of the heterogeneity discovered after adding $\Gamma$-distributed RAS can be attributed to the genetic code. One hidden class in all data sets examined appears to describe transition mutations exclusively, which occur in 2-fold degenerate third codon positions and, potentially, in first codon positions where only very similar amino acids can be accepted by selection. The other classes in the models describe other general forms of heterogeneity that were dependent on the data examined. Heterogeneity is often attributed to complex evolutionary events, such as molecular adaptation and environmental changes, but it appears that the simple structure of the genetic code alone could cause significant heterogeneity in sequence evolution.

Acknowledgments

The work presented here greatly benefited from discussion with many individuals during my visit to the Phylogenetics program at The Isaac Newton Institute for Mathematical Sciences, Cambridge, United Kingdom, during the winter of 2007. I would particularly like to thank Elizabeth Allman, John Rhodes, Peter Lockhart, Andrew Roger, Gavin Naylor, and Matthew Spencer. Joshua Herbeck kindly provided the groEL sequence alignments used in his study.

Literature Cited


Peter Lockhart, Associate Editor

Accepted May 11, 2008