Standard Codon Substitution Models
Overestimate Purifying Selection for Non-Stationary Data

Benjamin D. Kaehler¹, Von Bing Yap², and Gavin A. Huttley¹

¹Research School of Biology, Australian National University, Canberra, ACT, 2600, Australia
²Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117546, Singapore

ABSTRACT

Estimation of natural selection on protein-coding sequences is a key comparative genomics approach for de novo prediction of lineage specific adaptations. Selective pressure is measured on a per-gene basis by comparing the rate of non-synonymous substitutions to the rate of neutral evolution, typically assumed to be the rate of synonymous substitutions. All published codon substitution models have been time-reversible and thus assume that sequence composition does not change over time. We previously demonstrated that if time-reversible DNA substitution models are applied blindly in the presence of changing sequence composition, the number of substitutions is systematically biased towards overestimation. We extend these findings to the case of codon substitution models and further demonstrate that the ratio of non-synonymous to synonymous rates of substitution tends to be underestimated over three data sets of insects, mammals, and vertebrates. Our basis for comparison is a non-stationary codon substitution model that allows sequence composition to change. Model selection and model fit results demonstrate that our new model tends to fit the data better. Direct measurement of non-stationarity shows that bias in estimates of natural selection and genetic distance increases with the degree of violation of the stationarity assumption. Additionally, inferences drawn under time-reversible models are systematically affected by compositional divergence. As genomic sequences accumulate at an accelerating rate, the importance of accurate de novo estimation of natural selection increases. Our results establish that our new model provides a more robust perspective on this fundamental quantity.

Keywords: Codon models, natural selection, non-stationary, Markov model

INTRODUCTION

Understanding of the action of natural selection on protein-coding sequences underpins fundamental questions regarding evolution and has immediate practical ramifications. It has been known at least since the seminal contribution of King and Jukes (1969) that the operation of natural selection can be employed to identify functionally important sequence regions. Natural selection’s influence is now routinely estimated using probabilistic codon models of sequence evolution and reported in genome portals (Hubbard
et al., 2009). These models are widely employed for purposes as diverse as seeking to identify genes affecting the natural history of species (e.g. Chimpanzee Sequencing and Analysis Consortium, 2005) to aiding in the design of new materials (Maitip et al., 2015).

A reasonable model of codon evolution is critical to the measurement of selective pressure. Since the earliest such analysis that corrected for multiple substitutions at a single site (Kimura, 1977), it has been assumed that the process of evolution is time-reversible. While it is well established that this assumption is incorrect (e.g. Karlin and Ladunga, 1994), it is a modelling decision that is typically retained for analysis of biological sequences (nucleic or amino acid) for the sake of convenience. It has recently been shown that models of nucleotide substitution that do not make this assumption are statistically feasible in some instances where time-reversible models are not, and that assuming time-reversibility affects inference relating to genetic distance (Kaehler et al., 2015).

Pronounced differences in codon frequencies between species have been reported. For example, the codon AAA is approximately 250 times more abundant in the genome of Buchnera aphidicola than that of Streptomyces venezuelae (Nakamura et al., 2000). More generally it has been established that, within a genome, codon frequencies are strongly correlated with the nucleotide composition of flanking non-protein coding sequences (Muto and Osawa, 1987). The relationship between nucleotide composition and codon frequency has also been demonstrated to be highly predictive across bacteria, archaea and eukaryotes (Knight et al., 2001). As argued by Knight et al. (2001), these observations convincingly establish changes to neutral mutation processes as the primary driver for changes in codon frequency. Accordingly, accurate modelling of protein coding sequence evolution requires accommodating these properties of the underlying nucleotide substitution process.

The results of Yap et al. (2010) indicated that estimation of the neutral nucleotide substitution process is critical to robust estimation of the influence of natural selection. These authors introduced a codon model parameterisation that improved robustness in estimating the mode of natural selection by disentangling sequence composition and natural selection. These results depended on embedding the most general time-reversible nucleotide substitution process within the codon model. The resulting model further retained the property of time-reversibility and is thus also stationary.

In this paper we extend the techniques developed in Kaehler et al. (2015) to the context of codon substitution models, and reveal that the assumption of time-reversibility in codon substitution models introduces systematic bias.

Natural selection is frequently estimated using the parameter $\omega$, which allows the rate of synonymous and non-synonymous codon substitutions to vary from one another in a way that is not consistent with neutral evolution. Under a given model of neutral evolution, $\omega$ is defined so that $\omega = 1$ represents neutral evolution, $\omega > 1$ positive Darwinian selection, and $\omega < 1$ purifying selection. This quantity has a complicated history, with $\omega$ having also been called $dN/dS$ (Nei and Gojobori, 1986) and $K_a/K_s$ (Miyata and Yasunaga, 1980). Prior to the dominance of maximum-likelihood (subsequently Bayesian) methods, several empirical methods were devised for its calculation. The titles of two papers make apparent the need for simpler approaches at that time (Kimura, 1980; Nei and Gojobori, 1986).
Equivalently, \( \omega \) enters the parameterisation of a continuous-time Markov model of codon substitution as a multiplicative constant for all non-synonymous codon transitions (Muse and Gaut, 1994; Goldman and Yang, 1994) which is then fitted using maximum-likelihood or Bayesian techniques. The specifications for these models have developed over time but we focus on two choices. The first is that of Yang (1998), which we shall call Y98. This model is described in detail below but for the moment we will describe it as an extension of the nucleotide model of Hasegawa et al. (1985). It is currently the most widely used model of codon substitution. The second is one of those developed in Yap et al. (2010), which we shall denote CNFGTR, for conditional nucleotide frequency general time reversible. Y98, CNFGTR, and all previously published codon substitution models are time-reversible.

CNFGTR is an extension of the general time reversible (GTR) nucleotide substitution model (Lanave et al., 1984) that addresses a bias present in Y98 when nucleotides are not equiprobable. The properties of variants of several popular models, including a conditional nucleotide frequency extension of the model developed in Muse and Gaut (1994), were systematically explored in Yap et al. (2010). It was shown that CNFGTR most effectively reduced the above bias in Y98 and other biases shown to be present in the model of Muse and Gaut (1994).

Before continuing it is worth clarifying the relationship between non-stationarity and time-reversibility of Markov processes and models. In this context we call a substitution model time-reversible if all of the Markov processes that comprise the model are time-reversible. Similarly, non-stationary models are made up of non-stationary processes. The set of time-reversible models is nested in the set of stationary models, which is in turn nested in the set of non-stationary models. That is to say a time-reversible model cannot be non-stationary, but a non-stationary model can be time-reversible for certain choices of parameters.

In Kaehler et al. (2015) it was observed using parametric bootstraps (see Goldman, 1993) that for alignments of third codon position nucleotides from nuclear encoded genes in a triad of mammals, mitochondrial protein coding genes from the same mammals, and ribosomal RNA from microbes that time-reversible models never feasibly described the process that generated the data. A more general, non-stationary model could feasibly have generated the mammal nuclear encoded gene data set, and succeeded more often as a reasonable model for the mtDNA and rRNA data sets. It was also shown that the time-reversible models systematically overestimated genetic distance in a manner proportional to a non-parametric measure of non-stationarity. Further, departure from the molecular clock hypothesis was overstated by time-reversible models, demonstrating that biological inferences based on estimates of the number of substitutions were affected by use of the time-reversible models. This work required new theoretical insights into calculation of genetic distance in the non-stationary setting.

As an alternative to time-reversible codon substitution models we present the non-stationary GNC or general nucleotide codon model, which extends the non-stationary nucleotide model presented in Kaehler et al. (2015). We extend the theoretical insights in Kaehler et al. (2015) in this setting and show that it can be used to estimate \( \omega \) in a manner consistent with its forebears. We test large sets of alignments of mammals (human, mouse, opossum), vertebrates (human, xenopus, fugu), and ants (Florida carpenter ant, Indian jumping ant, Argentine ant) to show that GNC tends to fit the available data
better than time-reversible models. We further demonstrate that consistent with Kaehler et al. (2015), Y98 and CNFGTR tend to overestimate genetic distance in comparison with GNC in a manner proportional to a non-parametric measure of non-stationarity and link overestimation of genetic distance with underestimation of $\omega$. Similar to Yap et al. (2010), we use intronic data to show that Y98 estimates $\omega$ with a bias that is a function of sequence GC content, but that CNFGTR and GNC are not biased in this way. We also demonstrate that inference regarding the molecular clock hypothesis is affected by model time-reversibility.

The next section will specify GNC. The Results detail our empirical findings, which will be interpreted in the Discussion. The Material and Methods describe the provenance of our data, how we fitted the models, the time-reversible models for the purpose of comparison, and theoretical results demonstrating that $\omega$ is an appropriate measure of selective pressure under GNC.

NEW APPROACHES

The General Nucleotide Codon Model

The general nucleotide codon (GNC) model is a non-stationary, continuous-time Markov process. For time-reversible models, the location of the root of the phylogenetic tree has no bearing on the model’s predictions. This assumption is highly convenient, but incorrect. For non-stationary models the location of the root matters as the process evolves forward in time away from the root. Recall that the degree of a node is the number of edges that connect to that node. In this work we will make the distinction between trees that are node-rooted (eg. Fig. 1a), where the root node has degree three, and edge-rooted (eg. Fig. 1b), where the root node has degree two. All nodes other than the root are assumed to have degree three or one.

At the root of the tree, each codon is assigned an initial marginal probability with no further assumptions, so forming a 61-element row vector $\pi_0$. The conditional nucleotide processes for each codon position are identical except that any rate corresponding to a nucleotide substitution that results in a non-synonymous codon substitution is multiplied by the parameter $\omega$, with no further constraints. We will show in the Materials and Methods section that $\omega$ is equal to $K_a/K_s$ as defined in, for example, Goldman and Yang (1994).

Equivalently, we write that the codon process on an edge $(\mu, \nu)$ is defined by the Markov generator $Q^{\mu\nu}$ with off-diagonal elements labelled by codons $i_1i_2i_3$ and $j_1j_2j_3$:

$$q^{\mu\nu}_{i_1i_2i_3,j_1j_2j_3} = \begin{cases} 0, & \text{more than one } i_n \neq j_n, \\ r^{\mu\nu}_{i_n,j_n}, & \text{synonymous}, \\ \omega r^{\nu\mu}_{i_n,j_n}, & \text{non-synonymous}, \end{cases}$$

where $n \in \{1,2,3\}$, $\omega r^{\nu\mu}_{i_n,j_n}$ is the selective pressure on $(\mu, \nu)$, and $r^{\mu\nu}_{i_n,j_n}$ is in turn a matrix that defines the neutral nucleotide process. It is important to note that the node $\nu$ is further from the root node than $\mu$, which is the concrete consequence of each branch having a direction in time. Diagonal elements of $Q^{\mu\nu}$ are calculated in the usual fashion. Note that the marginal probabilities at node $\mu$, $\pi_\mu$, are absent from this formulation. We introduce a scale parameter $\tau^{\mu\nu}$ and scale $Q^{\mu\nu}$ such that $-\pi_0 \cdot \text{diag} Q^{\mu\nu} = 1$. The transition probability matrix on $(\mu, \nu)$ is then $P^{\mu\nu} = \exp\{Q^{\mu\nu} \tau^{\mu\nu}\}$. 

Unless we state otherwise, wherever we fit GNC we allow all parameters to vary by edge, so the rates are heterogeneous across lineages. In such cases each edge adds thirteen parameters to the model – twelve rate parameters and one selective pressure parameter. In some circumstances we will impose additional constraints.

At this point we digress briefly to outline the relationship between the scale parameter \( \tau \), time and genetic distance. Recall that genetic distance is measured as the expected number of substitutions between two nodes in the tree. Further, genetic distance is typically employed as a measure of “evolutionary time”. For time-reversible models, the expected number of substitutions along one edge equals the scale parameter. Accordingly, there is no confusion between what has been considered the time parameter and genetic distance and where we refer to a scale parameter in the context of a time-reversible model, it can safely be interpreted as the time parameter.

For non-stationary models, such as GNC, \( \tau \) does not necessarily equal the expected number of substitutions (Kaehler et al., 2015). Rather, the genetic distance can be calculated as

\[
d_{\text{GNC}}^{\mu \nu} = -\pi_{\mu} \int_{0}^{\tau_{\mu \nu}} \exp\{Q_{\mu \nu} s\} \, ds \, \text{diag} Q_{\mu \nu}.
\]

(1)

for the edge \((\mu, \nu)\). The derivation of this formula carries without modification from Kaehler et al. (2015). Genetic distance across multiple edges is the sum of the distances for those edges. Numerical methods for calculation of the integral of the matrix exponential are discussed in Kaehler et al. (2015).

RESULTS

GNC on a node-rooted tree is useful for comparison against time-reversible models

Our goal in this work is to assess how non-stationarity in the observed data affects inference drawn using time-reversible models. In particular we will compare GNC with Y98 from Yang (1998) and CNFGTR from Yap et al. (2010). GNC is non-stationary and was introduced in the New Approaches section, whereas CNFGTR and Y98 are time-reversible and are reviewed in the Materials and Methods section. In this context the phylogenetic tree topology is a nuisance parameter: we would like, as far as possible, to avoid questions regarding tree misspecification. It was shown in Chang (1996) under mild assumptions that a discrete-time model of character substitution could be consistently estimated using maximum likelihood on a node-rooted tree with not less than three leaves. All of the models considered in this work are submodels of the model in Chang (1996), but sadly the consistency proof does not automatically extend to continuous-time models, and we leave this as an open problem. However, in light of the results in Chang (1996), we do not wish to claim that GNC can be consistently estimated for two taxa. We therefore predominantly test the three models in question on triads of extant taxa using node-rooted trees (eg. Fig. 1a), for which the topology is unambiguous.

The use of a node-rooted tree is an approximation, as in reality we would have expected evolution to have proceeded on an edge-rooted tree. For instance, in Figure 1b, time flows differently in different parts of the branch connecting the opossum to the...
speciation event between mouse and human. It runs towards opossum in the lowermost part of the figure, but away from opossum in the part between the root and the mouse-human speciation event. For the reasons given above, we base our model on the tree in Figure 1a, in which the direction of time is wrongly specified along a part of this branch, but we hypothesise that a model incorporating this misspecification will still perform better than one which assumes time-reversibility throughout the tree. This hypothesis is based on work in Verbyla et al. (2013) and Kaehler et al. (2015), where similar assumptions do not seem to greatly affect model fit. Thus we hypothesise that when we later compare time-reversible models with GNC, GNC on a node-rooted tree will be a reasonable proxy for GNC on an edge-rooted tree. We tested this hypothesis using simulation.

As a basis for simulation we fitted GNC on an edge-rooted tree to ten protein-coding alignments of Human, Mouse, and Opossum. The alignments were of one-to-one orthologues of the Human genes ENSG00000074276, ENSG00000074621, ENSG00000090924, ENSG00000111641, ENSG00000138658, ENSG00000161647, ENSG00000164867, ENSG00000165813, and ENSG00000240303. The alignments were selected from a set of 4,150 alignments to maximise the observed difference in codon composition between the Human and Mouse sequences. Details of the data set are given in the Materials and Methods section. Difference in composition was measured using Jensen-Shannon divergence (JSD) (Lin, 1991). As we were concerned about the identifiability of GNC on an edge-rooted tree, we set all parameters excluding the scaling parameter to be equal on the Opossum branch and the internal branch. Otherwise all parameters were allowed to vary by branch. We rejected one alignment that resulted in a zero scaling parameter on the internal branch. We note that the goal here was not to fit a model for the purpose of inference, but rather to create a known, biologically plausible, edge-rooted model for the purpose of simulation.

We evaluated whether GNC on node-rooted trees could serve as a reliable proxy for the true non-stationary process under which the observed sequences evolved. To this end we simulated 100 alignments under each fitted, edge-rooted model. We then fitted GNC on a node-rooted tree, CNFGTR, and Y98 to each alignment. As in most of our experiments we allowed all parameters for GNC to vary by edge and allowed only the scale parameter and selective pressure parameter ($\omega$) to vary by edge for CNFGTR and Y98. We are interested in bias in estimated genetic distance and selective pressure. True bias was directly estimated as the generating model was known, and proxy bias was calculated by subtracting the GNC estimates of the parameters from the estimates under the time-reversible models. The results are shown in Figure 2.

We see that the bias introduced by fitting time-reversible models to data generated by non-stationary models dwarfs the bias introduced by ignoring the internal edge and fitting a model on a node-rooted tree rather than an edge-rooted tree, as most points fall close to the diagonal. We also note that these results show that our numerical model fitting procedure is capable of reconstructing the parameters of the generating models, at least for genetic distance and selective pressure.

There are two small artefacts in Figure 2, where the genetic distance bias is negative on the Human branch, and for selective pressure on the Opossum branch. We shall see that we are largely interested in overestimation, that is positive bias, of time-reversible models for genetic distance, so the small difference between the proxy bias and the true
bias on the Human branch are not material to our results. For the Opossum selective pressure results, we note that both proxy and true bias are close to zero.

The remainder of the results in this work concern models fitted to empirical data and on the basis of the simulation results in this section we will henceforth assume that the phylogenetic tree is node-rooted.

![Figure 1](image.png)

**Figure 1.** Tree topologies; (a) node-rooted, (b) edge-rooted. The stems illustrate the placement of the roots, but do not form part of the model. Arrows indicate the direction of time.

**The non-stationary model tends to fit better than the time-reversible models**

We performed two analyses to determine how well the non-stationary model GNC and the time-reversible models, Y98 and CNFGTR, fit empirical data. For each we fitted the models to large numbers of alignments of one-to-one orthologues separately. In all cases all GNC parameters were allowed to vary by branch, whereas only scale parameter and $\omega$ were allowed to vary by branch for Y98 and CNFGTR. All models were fitted using a node-rooted tree.

The first analysis was based on comparison of the Akaike Information Criterion (AIC) (Burnham and Anderson, 2002, pp. 60–64) on four data sets consisting of primate introns and protein-coding sequences. Details of the data sets are given in the Materials and Methods section. The second was an objective measure where the G-statistic (Sokal and Rohlf, 1995, pp. 686–697) of each fit was assessed using a simulated distribution generated using parametric bootstraps (Goldman, 1993). We performed this experiment on the three protein-coding alignments.

We make special mention here of the primate intron data set. Following the approach of Yap et al. (2010), this data set was used as representative of naturally occurring sequences evolving without the influence of natural selection for protein coding content and thus we expect $\omega = 1$. Also, the models were modified slightly when applied to the intronic data to allow stop codons.

Table 1 shows the results of the first experiment, where we used the AIC to determine which model best fitted each alignment. The number of parameters for each alignment were 67, 71, and 99 for Y98, CNFGTR, and GNC respectively. A detailed description of these parameters is given in the Materials and Methods section. Across the four data sets that we tested, GNC was identified as the best fitting model more often than both of the other models combined. The effect was most pronounced for the introns, which consist of much longer alignments than the other data sets (see Table 7).

In the second analysis, for each alignment, we tested the null hypothesis that the data were generated by the fitted model against the alternative hypothesis that they were...
not. The use of parametric bootstraps of the G-statistic for this purpose is detailed in the Materials and Methods section. We used 49 bootstrap iterations for every test. This number is low for this type of test, so it is worth mentioning that these calculations were particularly computationally intensive, requiring the equivalent of well over 100,000 hours of computation on a single computing core. Table 2 shows the proportion of alignments in each data set that rejected the null hypothesis at nominal 5% significance. While none of the models or data sets reproduced the results in Kaehler et al. (2015), where rejection rates were as low as 6%, GNC consistently performed better than CNFGTR, which in turn outperformed Y98. The full distributions of the p-values for these tests are shown in Figure 3.

**Time-reversible models overestimate genetic distance in a manner proportional to observed non-stationarity**

It was observed in Kaehler et al. (2015) that time-reversible nucleotide substitution models tend to systematically overestimate genetic distance in comparison with a non-stationary model, in a manner proportional to observed non-stationarity. We report the same phenomenon for codon substitution models. We fitted GNC, CNFGTR, and Y98 to the mammal, vertebrate, and ant data sets using node-rooted trees. All GNC parameters were allowed to vary by branch, whereas only the scale parameter and selective pressure parameter (ω) were allowed to vary by branch for Y98 and CNFGTR.
Table 1. Model with minimum AIC by data set. General Nuclear Codon (GNC) model is non-stationary. Conditional Nucleotide Frequency GTR (Yap et al., 2010, CNFGTR) and Yang (1998, Y98) are time-reversible. Data sets consisted of 4,150 alignments of mammal protein-coding sequences, 2,019 protein-coding ant alignments, 2,907 protein-coding alignments of vertebrates, and 13,046 primate intronic alignments.

<table>
<thead>
<tr>
<th></th>
<th>GNC</th>
<th>CNFGTR</th>
<th>Y98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammals</td>
<td>59.9%</td>
<td>37.7%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Ants</td>
<td>79.2%</td>
<td>16.3%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>55.2%</td>
<td>36.4%</td>
<td>8.4%</td>
</tr>
<tr>
<td>Primate Introns</td>
<td>91.4%</td>
<td>8.1%</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

Table 2. Percentage of alignments for which the fits rejected the null hypothesis that the data was feasibly generated by the fitted model at nominal 5% significance. Hypotheses were tested using the G-statistic and 49 parametric bootstrap iterations. Data sets consisted of 4,150 alignments of mammal protein-coding sequences, 2,019 protein-coding ant alignments, and 2,907 protein-coding alignments of vertebrates.

<table>
<thead>
<tr>
<th></th>
<th>GNC</th>
<th>CNFGTR</th>
<th>Y98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammals</td>
<td>80.4%</td>
<td>83.2%</td>
<td>95.9%</td>
</tr>
<tr>
<td>Ants</td>
<td>74.2%</td>
<td>85.8%</td>
<td>95.5%</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>86.9%</td>
<td>87.2%</td>
<td>89.0%</td>
</tr>
</tbody>
</table>

Define $d_{Y98}$ and $d_{CNFGTR}$ as the expected number of substitutions over a path through a tree as inferred from the Y98 and CNFGTR models respectively. $d_{GNC}$ is defined above in (1). As justified by our simulation and model fit results, we quantified the genetic distance bias of time-reversible models using $d_{Y98} - d_{GNC}$ and $d_{CNFGTR} - d_{GNC}$. For a pair of sequences, we take conservation of codon composition as a measure of stationarity. We use the Jensen-Shannon divergence (JSD) (Lin, 1991) for this purpose. Figure 4 shows the relationship between JSD and genetic distance bias for three data sets. Each point in those scatter plots represents a single alignment, where the JSD and genetic distance are taken for the pair of taxa with maximal JSD within the triad. Alignments were excluded from the analysis if their scale parameter on any branch attained a preset maximum value (of 10). (See the New Approaches section for a discussion of the scale parameter.) Table 3 shows how many alignments were eliminated under this criteria for each model and data set. Table 4 summarises the slopes and intercepts of the median regression lines for Figure 4.

The overwhelming observation is that genetic distance bias tends to be positive, and increases with increasing non-stationarity. For the mammal and ant data sets, Y98 yields genetic distance biases that grow more quickly with increasing JSD than those for CNFGTR. For these data sets the intercepts are close to zero. For the vertebrate data set the same trend is present but less clear. We observe that for this data set the fitted scale parameters for the time-reversible models were often at the preset bound of 10 (see Table 3), indicating that to these models the number of substitutions often looks...
<table>
<thead>
<tr>
<th></th>
<th>GNC</th>
<th>CNFGTR</th>
<th>Y98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ants</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mammals</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>104</td>
<td>1281</td>
<td>1832</td>
</tr>
</tbody>
</table>

**Table 3.** Number of fits that were excluded because a branch scale parameter attained the maximum allowed value (10) by model and data set.

<table>
<thead>
<tr>
<th></th>
<th>Y98</th>
<th>CNFGTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td>Intercept</td>
<td>Slope</td>
</tr>
<tr>
<td>Ants</td>
<td>6.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Mammals</td>
<td>7.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>12.1</td>
<td>3.6</td>
</tr>
</tbody>
</table>

**Table 4.** Median regression for genetic distance bias against Jensen-Shannon Divergence (JSD). Genetic distance biases are $d_{Y98} - d_{GN}$ or $d_{CNFGTR} - d_{GN}$, where $d$ is the genetic distance under the model indicated by the subscript. Genetic distance is the expected number of substitutions. Measurement was between the pair of taxa with maximal JSD for a given alignment. Mammal, ant, and vertebrate data sets contained 4,140 and 4,149 (Y98 and CNFGTR), 2,019 (both), and 1,074 and 1,622 (Y98 and CNFGTR) alignments respectively.

...saturated. We speculate that this is a source of noise that obscures these results, even if the branch lengths are not completely saturated.

**Time-reversible models underestimate $\omega$ in a manner proportional to their overestimation of genetic distance**

A key feature of a codon model is its ability to estimate natural selection. We will show that the parameter representing natural selection, $\omega$, as estimated using time-reversible models, is systematically biased in comparison to its measurement using our non-stationary model. For this purpose we fitted GNC, CNFGTR, and Y98 to the mammal, vertebrate, and ant data sets using node-rooted trees. All GNC parameters were allowed to vary by branch, whereas only the scale parameter and selective pressure parameter ($\omega$) were allowed to vary by branch for Y98 and CNFGTR.

We expect that if branch length is overestimated, then the corresponding parameter $\omega$ should be underestimated to compensate. We reason that to obtain an observed level of amino acid conservation, if the genetic distance is greater, then the damping effect of $\omega$ must also be greater. As we have estimated branch length and $\omega$ separately for each branch we can compare genetic distance bias and $\omega$ bias directly, as we do in Figure 5. Genetic distance bias is calculated as described in the last section and $\omega$ bias is calculated analogously as $\omega_{GNC} - \omega_{Y98}$ and $\omega_{GNC} - \omega_{CNFGTR}$. Again, where the branch scale parameter attained its maximum allowed value for any branch in an alignment, that alignment was excluded from the analysis (see Table 3).

We observed a strong negative correlation between genetic distance bias and $\omega$ bias.
As lower $\omega$ means that non-synonymous substitutions are more harshly penalised, then underestimating $\omega$ corresponds to overestimating purifying natural selection. While the relationship between genetic distance bias and $\omega$ bias is noisy, we note that the majority of the mass for each scatter plot in Figure 5 falls in the second and fourth quadrants. For this reason we do not speculate on the form of the functional relationship, but draw regressions through the origin for each plot.

**Biases that affect time-reversible models do not affect a simple non-stationary model**

Yap et al. (2010) used experiments on intronic regions as controls to test for bias in estimates of $\omega$ in the time-reversible codon models that they considered. This method has great appeal because it allows us to test our models of natural selection on naturally occurring sequences that we expect to have evolved under the neutral process for protein coding content, and as such constitute a negative control for which we would expect $\omega = 1$. The models used for these experiments had to be modified slightly to allow stop codons in a similar fashion to Yap et al. (2010).

As Yap et al. (2010) compared estimates of $\omega$ with the measured GC content of the sequences, we do the same in Figure 6, with details in Table 5. In each case we fitted GNC, CNFGTR, and Y98 to 13,046 alignments of marmoset, macaque, and human, with a median length of 11,130 nt.

We verify the finding in Yap et al. (2010) that $\omega_{Y98}$ has a strong bias that varies with GC content. Lower GC content biases $\omega_{Y98}$ toward underestimation and higher GC content biases $\omega_{Y98}$ toward overestimation. Overall, median $\omega_{Y98}$ is consistently less than one. It is interesting to note that of the three branches, the median estimate of $\omega$ is close to one for all three models for the marmoset branch. The true root of the evolutionary tree for these species is expected to lie on this edge, so we speculate that this is a manifestation of a different underlying process operating on this edge.

The key finding for our tests on intronic data is that $\omega$ bias does not appear to vary by GC content for GNC or CNFGTR, with the slopes for the median regression lines for these models being close to zero. By inspecting the median $\omega_{CNFGTR}$ we see that it again slightly underestimates $\omega$ for the ingroup edges. This effect is not pronounced for the GNC results.

That Y98 produces $\omega$ estimates that are biased when the nucleotides are not equiprobable is a consequence of the presence of codon probability parameters in the Markov generator (see Materials and Methods), so that substitution rates are confounded with codon probability parameters. A full description of this phenomenon is given in Lindsay et al. (2008). Models such as that of Muse and Gaut (1994), where the transition rates are multiplied by nucleotide frequencies rather than codon frequencies are less susceptible to this type of bias, and this property is extended by design to conditional nucleotide frequency models such as CNFGTR (Yap et al., 2010). The Muse and Gaut (1994) model has other biases that are introduced because codon probabilities do not appear to be decomposable into nucleotide probabilities, which are also addressed by CNFGTR. GNC is not exposed to bias in the same way because the codon probability parameters do not enter the formulation of the Markov generator, as explained in the New Approaches section.
Table 5. $\omega$ is strongly biased when using Y98, weakly biased when using CNFGTR, and least biased using GNC. $\omega = 1$ is expected for these 13,046 intronic alignments of human, macaque, and marmoset. Table shows median $\omega$ and the slope of the median regression of $\omega$ versus GC content, as shown in Figure 6.

<table>
<thead>
<tr>
<th></th>
<th>Median $\omega$</th>
<th>Median Regression slope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GNC</td>
<td>CNFGTR</td>
</tr>
<tr>
<td>Marmoset</td>
<td>1.003</td>
<td>1.006</td>
</tr>
<tr>
<td>Macaque</td>
<td>0.995</td>
<td>0.977</td>
</tr>
<tr>
<td>Human</td>
<td>0.991</td>
<td>0.975</td>
</tr>
</tbody>
</table>

Non-stationary and time-reversible models imply different conclusions regarding the molecular clock

To show that inferences drawn from the assumption of a non-stationary codon model differ to those from time-reversible models, we investigated how often the molecular clock hypothesis was violated for the mammals data set. We also performed this experiment to test the generality of results obtained in Kaehler et al. (2015) regarding the effect of using time-reversible models to test the molecular clock hypothesis. The molecular clock hypothesis in this context asserts that the genetic distance along the mouse and human branches should be identical. We expected that the molecular clock test would be rejected in the majority of cases (eg. Huttley et al., 2007).

We tested the hypothesis in two experiments where likelihood ratio tests (LRTs) were performed on 4,150 alignments. These tests are variants of the likelihood form of the relative rate tests established in Muse and Weir (1992). In all cases parameters other than the scale parameter and $\omega$ were constrained to be equal for all branches for the time-reversible models and allowed to vary by lineage for GNC and models were fitted on a node-rooted tree. In both experiments the genetic distance was also constrained to be equal on the human and mouse branches under the null hypothesis, and allowed to vary under the alternative. For the first experiment, $\omega$ was allowed to vary by branch. For the second, $\omega$ was constrained to be equal across all branches for the time-reversible models. The slightly unusual set-up for the latter was to enable comparison with the analogous test performed for Kaehler et al. (2015), which was performed without the complicating factor of including a selective pressure parameter.

The results are summarised in Table 6, which shows the proportion of alignments where the null hypothesis was rejected at nominal 5% significance, and Figure 7, which shows cumulative distribution functions for the LRT p-values. Where $\omega$ was allowed to vary by branch, the non-stationary model rejected the null hypothesis more often than the time-reversible models. Where $\omega$ was constrained for the time-reversible models, the opposite result was observed, confirming the result in Kaehler et al. (2015). In both cases, the results for the two time-reversible models were almost indistinguishable.
Table 6. Proportion of alignments for which the molecular clock was rejected at nominal 5% significance for 4,150 alignments of human, mouse, and opossum protein coding genes. For the first experiment, $\omega$ was allowed to vary by branch in every case, and for the second $\omega$ is constrained to be equal across all branches for the time-reversible models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\omega$ Varies</th>
<th>$\omega$ Constrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNC</td>
<td>67.7%</td>
<td>67.7%</td>
</tr>
<tr>
<td>CNFGTR</td>
<td>66.8%</td>
<td>72.7%</td>
</tr>
<tr>
<td>Y98</td>
<td>67.0%</td>
<td>73.1%</td>
</tr>
</tbody>
</table>

**Non-stationary and time-reversible models imply different conclusions regarding whether $\omega$ is constant between lineages**

As $\omega$ can be used to examine the difference between selection on different branches (eg. Yang, 1998; Huttley et al., 2000), we also tested how inference regarding $\omega$ on the human and mouse branches of the mammal data was affected by model choice.

The null hypothesis was that $\omega$ was equal between the human and mouse branches, against the alternative that it was not. We again performed LRTs on this basis. In all instances no genetic distances were constrained to be equal. As for our other experiments, parameters other than $\omega$ and the scale parameter were equal across the tree for Y98 and CNFGTR and allowed to vary by branch for GNC and all models were fitted on a node-rooted tree. The results for the 4,150 mammals alignments are shown in Figure 8. Similar to the first clock test, which most closely resembles this experiment, GNC indicated more violations of the null hypothesis, rejecting the null hypothesis at nominal 5% significance in 26.7% of cases. Again, the two time-reversible models gave almost identical results, both rejecting the null in 16.6%.

**DISCUSSION**

We applied simulations, tests of model fit, and comparison of statistical inference to examine some ways the assumption of time reversibility might mislead the scientist. Two representative time-reversible models were chosen for comparison (Y98 and CNFTGR), and a new, simple non-stationary model was developed (GNC). We tested three large sets of protein-coding genetic sequences from mammals, ants, and vertebrates, and one large control set of intronic primate data.

We firstly established via simulation under the assumption that phylogenetic trees are edge-rooted that, for the purposes of comparing Y98 and CNFTGR with GNC, it was acceptable to fit models on node-rooted trees. Our simulations also showed that the parameter-rich GNC model was able to successfully reconstruct the parameters in which we were interested with the amount of data that we had at hand.

In our tests of model fit, for each data set, GNC had the best AIC score of the three models tested in the majority of cases. This effect was most pronounced for the ant and intron data sets, where GNC was selected greater than 79% and 91% of the time. We reason that it is safe to assume that non-stationarity is present in all of our data sets.

We included an objective test of model fit that is capable of rejecting model fits if the
model could not feasibly have generated the data. Again, GNC performed the best of the three models. As an aside, the ordering was consistent for every test we performed: GNC always tended to fit better than CNFGTR, which always outperformed Y98. In Kaehler et al. (2015), the equivalent experiment showed that a non-stationary nucleotide model was rejected at close to the type I error rate for one data set. Here the rejection rates were much higher (> 74% for GNC in all cases). This shows that it is unlikely that any of the models considered here could feasibly have generated any of our data sets. The difference between GNC and the non-stationary nucleotide model considered in Kaehler et al. (2015) is that the nucleotide model was the most general possible time-homogeneous (on branches), continuous-time Markov process. The analogous codon model would have 5,550 parameters for a triad of taxa, so we are forced to make more modelling assumptions in this context.

We conclude that of the models considered, GNC tends to fit the data best, but that it may be possible to improve on GNC.

For every data set, genetic distances estimated under both time-reversible models were overestimated in comparison to those obtained from the non-stationary model. For the ant and mammal data sets it was clear that this overestimation was proportional to a non-parametric measurement of non-stationarity, meaning that if the data supported time-reversibility, GNC tended to recapitulate the results obtained from CNFGTR or Y98. For the vertebrates the signal was less clear, although there are some indications that this data set exists at the limits of our inference. For the ants and mammals, this overestimation grew more rapidly with non-stationarity for Y98 than CNFGTR.

In conjunction with the results in Kaehler et al. (2015), we observe that overestimation of genetic distance by time-reversible models as proportional to non-stationarity is a consistent trend over data sets from vertebrates, invertebrates, and microbes for four different time-reversible nucleotide and codon models. Why fitting time-reversible models to data generated by non-stationary processes should lead to this phenomenon is an open question.

Next we addressed how inference about natural selection might be affected by the assumption of time-reversibility. We hypothesised that \( \omega \) would be partially confounded with genetic distance, and found strong evidence for both time-reversible models and all data sets that when genetic distance was overestimated, \( \omega \) tended to be underestimated. As overestimation of genetic distance is reduced when less non-stationarity is observed, so too is underestimation of \( \omega \), so again nothing is lost by using GNC rather than CNFGTR or Y98. Again, we observe that the effect is more pronounced in Y98 than CNFGTR, particularly for the ants.

In Yap et al. (2010), the conditional nucleotide frequency models of which CNFGTR is an example were introduced to combat bias that is observed in Y98 and other time-reversible codon models. We confirmed that GNC is also immune to such biases by conducting control experiments on intronic data from three primates. We perceive this advantage is a result of the simple form of the GNC parameterisation. The expected result of the experiment was that \( \omega \) should be approximately equal to one. This is what we observed for GNC, with Y98 again systematically underestimating \( \omega \) in addition to being biased where GC content was not equal to 0.5. CNFGTR is interesting because it only slightly underestimates \( \omega \) (with median values of roughly 1.01, 0.98, and 0.98 for the three branches), whereas approximately 91% of the intron data set selects GNC via
its AIC score, indicating the presence of non-stationarity. This is in keeping with our other results if the amount of non-stationarity is low, as in that case we would expect $\omega$ to be only slightly underestimated. It is likely that as the intron alignments are quite long (median length 11,130 nt), the power of the AIC test was sufficient to detect this small amount of non-stationarity.

To demonstrate that assuming time-reversibility affects inference regarding key phylogenetic questions, we performed a molecular clock test on the mammal data set, using LRTs to test the hypothesis that the genetic distances along the human and mouse branches are equal. For the first experiment we allowed $\omega$ to vary by branch, in keeping with our other experiments. We found that the two time-reversible models, regardless of their apparent difference in specification and other experimental results here presented, gave almost indistinguishable results for this test. The non-stationary model, however, rejected the clock hypothesis more often. Then, to facilitate comparison with the results in Kaehler et al. (2015), we held $\omega$ constant for each branch only for the time-reversible models, and repeated the experiment. As a result the time-reversible models became substantially more likely to reject the molecular clock and more so than GNC, bringing the results into agreement with those in Kaehler et al. (2015). We speculate that as $\omega$ and genetic distance are partially confounded, $\omega$ was able to compensate for varying genetic distance in the first experiment. It is interesting, then that GNC, which has considerably more degrees of freedom, was more likely to reject the null in the first experiment, pointing to deeper differences between the non-stationary and time-reversible models.

Finally, we tested the hypothesis that $\omega$ is equal for the human and mouse branches in the mammal data. In keeping with the first molecular clock test, which it most closely resembled as genetic distance was allowed to vary by branch, GNC was found to be more likely to reject the hypothesis that $\omega$ was equal and Y98 and CNFGTR were found to be almost identical. Again, time-reversibility seems to be the key distinction between the performance of models in this test.

Our tests of equality across lineages of genetic distance and natural selection show that these quantities clearly vary by lineage. It is therefore natural to assume that non-stationarity must vary from branch to branch as well. We have shown that the bias introduced by fitting time-reversible models to non-stationary data varies with non-stationarity. These biases cannot be a simple scaling of a phylogenetic tree, and so we speculate that inferences based on time-reversible models outside those analysed in this work must also be susceptible to bias.

Understanding the extent to which the historical operation of natural selection has shaped the distribution of genetic variation has, to a very large extent, derived from application of codon models of sequence evolution. In the simplest sense, our ability to draw inference relies on how well these models represent the process of neutral sequence evolution. As we demonstrated previously (Kaehler et al., 2015), utilising time-reversible nucleotide substitution processes distorts our estimation of the number of events in a manner that is proportional to the extent of non-stationarity. Non-stationarity is common across the tree of life (e.g. Karlin and Ladunga, 1994). In this work, we have shown that the biases that were evident in time-reversible nucleotide models manifest in the codon case in such a manner as to underestimate the ratio of non-synonymous to synonymous substitutions. As a consequence, the application of the time-reversible models to sequences with deep evolutionary divergence seems likely to give rise to
estimates of natural selection that are artefactual, bringing into question conclusions regarding historical shifts in the operation of selection as have been recently reported (Sunagar and Moran, 2015).

**MATERIALS AND METHODS**

**Data**

We obtained the mammal and vertebrate data sets by downloading protein coding sequences from Ensembl 68. Each multiple sequence alignment in the mammal data set was made up of three one-to-one orthologues from *Homo sapiens, Mus musculus*, and *Monodelphis domestica* and each alignment in the vertebrate data set consisted of three one-to-one orthologues from *Homo sapiens, Takifugu rubripes*, and *Xenopus tropicalis*. PyCogent (Knight et al., 2007) was used to download these sequences.

Alignments of intronic one-to-one orthologous sequences for *Homo sapiens, Macaca mulatta*, and *Callithrix jacchus* were also downloaded from Ensembl 81. Regions annotated in Ensembl as low complexity were masked and those columns removed from the alignments in a manner that preserved naturally occurring trinucleotides. Again, PyCogent was used to obtain the data.

Another set of protein coding sequences were downloaded from the Ant Genome Portal (Munoz-Torres et al., 2011) for *Camponotus floridanus, Harpegnathos saltator*, and *Linepithema humile*. One-to-one orthologues sequences were selected using three-way reciprocal BLAST (Camacho et al., 2009), meaning that for each alignment, every sequence was the top BLAST hit for the other two sequences in both directions.

All protein coding sequences were aligned using PyCogent assuming a CNFHKY model (Yap et al., 2010).

In every instance any column in a multiple sequence alignment that contained a non-nucleotide character was excluded and only alignments with at least 500 remaining codons were analysed.

Some basic statistics for the four data sets are given in Table 7.

**Statistical methods and algorithmic implementation**

All substitution models were fitted using PyCogent (Knight et al., 2007). GNC has not yet been incorporated into the PyCogent library, but is available with other ancillary code at [https://bitbucket.org/nonstationary/codon](https://bitbucket.org/nonstationary/codon). All simulations were also performed using PyCogent.

As stated above, Y98 is the codon model specified in Yang (1998). In this model,
a 61-element stationary marginal probability distribution \( \pi \) is defined on the space of all amino acid coding codons. The Markov generator, labelled by codons \( i \) and \( j \), has off-diagonal elements

\[
q_{ij} = \begin{cases} 
\pi_j, & \text{one synonymous transversion,} \\
\kappa \pi_j, & \text{one synonymous transition,} \\
\omega \pi_j, & \text{one non-synonymous transversion,} \\
\omega \kappa \pi_j, & \text{one non-synonymous transition,} \\
0, & \text{otherwise.}
\end{cases}
\]

\( Y98 \) is time-reversible.

The CNFGTR model is defined in Yap et al. (2010). It also has a 61-element stationary marginal codon probability distribution. Its Markov generator’s off-diagonal elements, labelled by codons \( i_{123} \) and \( j_{123} \), is given by

\[
q_{i_{123}j_{123}} = \begin{cases} 
0, & \text{more than one } i_n \neq j_n, \\
r_{i_n,j_n} \pi_{j_n\{j_m\}_{m \neq n}}, & \text{synonymous, } i_n \neq j_n, \\
or_{i_n,j_n} \pi_{j_n\{j_m\}_{m \neq n}}, & \text{non-synonymous, } i_n \neq j_n,
\end{cases}
\]

where \( n, m \in \{1, 2, 3\} \), \( r_{i,j} \) are the elements of a \( 4 \times 4 \) symmetric matrix, and \( \pi_{j_n\{j_m\}_{m \neq n}} \) is the probability of observing \( j_n \) given the other two codon positions \( j_m \). CNFGTR is time-reversible.

The formulation of the above models and GNC with 61 states is, of course, dependent on the genetic code associated with the alignments to which the models are fitted. The software that we provide is agnostic to genetic code and is easily configurable for alternatives, as we did for the tests on intronic data where we allowed stop codons.

In every case \( \omega \) represents the influence of natural selection, which is sometimes also referred to as \( dN/dS \) (eg. Yap et al., 2010) or \( K_a/K_s \) (eg. Goldman and Yang, 1994).

Except for our simulation studies, we fitted the models to alignments of three protein-coding DNA sequences assuming a node-rooted tree (Fig. 1a). For the time-reversible models, the scale parameter \( \tau \) and selective pressure \( \omega \) were allowed to vary by branch, except where otherwise noted. For the non-stationary model, all parameters were usually allowed to vary on each branch. In every case, marginal codon probabilities were fitted along with the other parameters.

In this standard scenario, the initial marginal codon probabilities contributed 60 parameters to each model. For \( Y98 \), \( \kappa \) contributes one additional parameter and \( \tau \) and \( \omega \) contribute two parameters per branch, making a total of 67 parameters for the three-taxon tree. For CNFGTR, the GTR parameters added five parameters to the model and \( \tau \) and \( \omega \) again added two parameters per branch, summing to 71 parameters. GNC was most parameter-rich, with 12 rate parameters plus \( \omega \) per branch for 99 parameters overall.

Genetic distances were calculated as the expected number of substitutions per codon. The expected number of substitutions for non-stationary models was calculated following the method of Kaehler et al. (2015), as noted in the New Approaches section.

For convenience, we here reproduce from Kaehler et al. (2015) our description of how parametric bootstraps (Goldman, 1993) and the G-statistic (Sokal and Rohlf, 1995,
pp. 686–697) were used to calculate p-values that objectively reflect goodness of fit for a particular model. The null hypothesis says that the alignment is generated by the fitted model with parameter values set at our estimates. The expected site-pattern counts under the model is thus the probability of the pattern multiplied by the alignment length. The alternative is the unrestricted multinomial model, as described in Goldman (1993), taken as the observed site-pattern counts in the alignment. The G-statistic is computed from the expected and observed counts using the conventional expression (Goldman, 1993, pp. 686–697). The bootstrap procedure is to

1. simulate 49 alignments of the same length as the fitted alignment under the null hypothesis;
2. perform the original fit on each simulated alignment; and
3. calculate the proportion of fitted G statistics that exceed or equal that of the original statistic.

The result is the G-statistic parametric bootstrap p-value.

**Estimation of selective pressure under non-stationary models**

We show that the definition of \(K_a/K_s\) used in Goldman and Yang (1994) transfers unscathed to the context of non-stationary Markov models. Further, it is equivalent to the parameter \(\omega\) in the following standard definition (see, for example, Yang, 2006, p. 59)

Define the instantaneous transition rates of a codon substitution process as

\[
q_{ij} = \begin{cases} 
0, & \text{more than one difference,} \\
r_{ij}, & \text{one synonymous difference} \\
\omega r_{ij}, & \text{one non-synonymous difference}
\end{cases}
\]

where \(r_{ij}\) is some representation of an evolutionarily neutral process and \(\omega\) summarises the influence of natural selection. Note that \(r_{ij}\) is allowed to vary by transition in this formulation but has traditionally been further constrained to enforce, for instance, time-reversibility of the process. We denote the marginal codon probability row vector \(\pi(t)\) for time \(t > 0\).

Define two new transition matrices:

\[
q_{ij}^S = \begin{cases} 
0, & \text{one synonymous difference,} \\
r_{ij}, & \text{otherwise,}
\end{cases}
\]

and

\[
q_{ij}^N = \begin{cases} 
0, & \text{one non-synonymous difference,} \\
r_{ij}, & \text{otherwise,}
\end{cases}
\]
so that $Q = \omega Q^N + Q^S$, where $Q$, $Q^N$, and $Q^S$ are the matrices comprised of $q_{ij}$, $q^N_{ij}$, and $q^S_{ij}$, respectively, for their off-diagonal elements. Diagonal elements are calculated in the usual way for Markov generators.

Following Goldman and Yang (1994), we define $\rho_s$ and $\rho_a$ to be the rates of synonymous and non-synonymous substitution per codon. They can be shown to be

$$
\rho_s = -\pi(t) \text{diag } Q^S \quad \text{and} \quad \rho_a = -\omega \pi(t) \text{diag } Q^N.
$$

Further, we define $\rho^1_s$ and $\rho^1_a$ as the potentials of synonymous and non-synonymous substitutions when no selective constraints exist at the amino acid level, that is $\rho^1_s = -\pi(t) \text{diag } Q^S$ and $\rho^1_a = -\pi(t) \text{diag } Q^N$. These parameters correspond to $\rho^\infty_s$ and $\rho^\infty_a$ in Goldman and Yang (1994). Likewise, $3\rho^1_s$ and $3\rho^1_a$ represent the potentials per codon. As these have historically been termed the numbers of synonymous and non-synonymous nucleotide sites per codon, we continue that terminology. The numbers of synonymous substitutions per synonymous site and non-synonymous substitutions per non-synonymous site are $K_s = \rho_s / 3\rho^1_s = 1/3$ and $K_a = \rho_a / 3\rho^1_a = \omega / 3$. It is therefore clear that

$$
\omega = \frac{K_a}{K_s},
$$

which shows that the usual definition and understanding of $\omega$ is sufficient in the context of non-stationary Markov models.

ACKNOWLEDGMENTS

We acknowledge funding support from the Australian Research Council and the National Health and Medical Research Council awarded to G. A. H. and V. B. Y., and from the Singapore Ministry of Education (AcRF Grant R-155-000-127-112) awarded to V.B.Y.

REFERENCES


Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast


Figure 3. Empirical cumulative distribution functions of p-values for the null hypothesis that the alignment could have been generated by the fitted model. Hypotheses were tested using the G-statistic and 49 parametric bootstrap iterations. Datasets consisted of 4,150 alignments of mammal protein-coding sequences (a), 2,019 protein-coding ant alignments (b), and 2,907 protein-coding alignments of vertebrates (c).
Figure 4. The genetic distance bias increases with Jensen-Shannon Divergence (JSD). Genetic distance biases are $d_{Y98} - d_{GNC}$ or $d_{CNFGTR} - d_{GNC}$, where $d$ is the genetic distance under the model indicated by the subscript. Genetic distance is the expected number of substitutions. Measurement was between the pair of taxa with maximal JSD for a given alignment. In every case the time-reversible models tend overwhelmingly towards overestimation. Solid lines show quantile regressions for 25%, 50%, and 75% quantiles. (a) shows 4,140 (Y98) and 4,149 (CNFGTR) alignments of human, mouse, and opossum protein coding genes. (b) shows 2,019 alignments of Florida carpenter ant, Argentine ant, and Indian jumping ant. (c) shows 1074 (Y98) and 1,622 (CNFGTR) alignments of human, xenopus, and fugu. Plots have been cropped to remove outliers.
ω bias is negatively correlated with genetic distance bias. An empirical relationship exists between genetic distance bias and ω bias. In every case the overestimation of genetic distance is linked to underestimation of ω. Solid lines show linear regressions through the origin. (a) shows 4,140 (Y98) and 4,149 (CNFGTR) alignments of human, mouse, and opossum protein coding genes. (b) shows 2,019 alignments of Florida carpenter ant, Argentine ant, and Indian jumping ant. (c) shows 1,074 (Y98) and 1,622 (CNFGTR) alignments of human, xenopus, and fugu. Plots have been cropped to remove outliers.
Figure 6. $\omega$ is systematically biased when estimated using Y98 and CNFGTR, but not GNC. $\omega = 1$ is expected for these 13,046 intronic alignments of human, macaque, and marmoset. Scatter plots relate GC content to estimates of $\omega$. Solid lines show quantile regressions for 25%, 50%, and 75% quantiles.
Figure 7. Molecular clock inference is affected by model time-reversibility and assumptions regarding $\omega$. Empirical cumulative distribution functions of likelihood ratio test p-values between constrained clock-like and unconstrained models based on Y98, CNFGTR, and GNC models over 4,150 alignments of human, mouse, and opossum protein coding genes. For (i), $\omega$ was allowed to vary by branch in every case; for (ii) $\omega$ was constrained to a single value for the tree for the time-reversible models.

Figure 8. Empirical cumulative distribution functions of likelihood ratio test p-values between models with equal $\omega$ for mouse and human branches and unconstrained models based on Y98, CNFGTR, and GNC models over 4,150 alignments of human, mouse, and opossum protein coding genes.