Is the General Time-Reversible Model Bad for Molecular Phylogenetics?

Jeremy G. Sumner1,*, Peter D. Jarvis1, Jesús Fernández-Sánchez2, Bodie T. Kaine1, Michael D. Woodhams3, and Barbara R. Holland1

1School of Mathematics and Physics, University of Tasmania, Hobart 7001, Australia; and 2Department de Matemática Aplicada, Universitat Politècnica de Catalunya, Barcelona 08028, Spain; *Correspondence to be sent to: School of Mathematics and Physics, University of Tasmania, Private Bag 37, Hobart, TAS, 7001, Australia E-mail: jsumner@utas.edu.au.

Received 3 November 2011; reviews returned 9 January 2012; accepted 19 March 2012

Associate Editor: Edward Susko

The general time-reversible (GTR) model (Tavaré 1986) has been the workhorse of molecular phylogenetics for the last decade. GTR sits at the top of the ModelTest hierarchy of models (Posada and Crandall 1998) and, usually with the addition of invariant sites and a gamma distribution of rates across sites, is currently by far the most commonly selected model for phylogenetic inference (see Table 1).

However, a recent publication (Sumner et al. 2012) shows that GTR, along with several other commonly used models, has an undesirable mathematical property that may be a cause of concern for the thoughtful phylogeneticist. In mathematical terms, the problem is simple: matrix multiplication of two GTR substitution matrices does not return another GTR matrix. It is the purpose of this article to give examples that demonstrate why this lack of closure may pose a problem for phylogenetic analysis and thus add GTR to the growing list of factors that are known to cause model misspecification in phylogenetics.

The phylogenetics literature is rich with other examples of model misspecification, each with the potential to cause problems for inference. For example, Galtier (2004) considers the effect of nonindependence of sites, whereas Galtier and Gouy (1995) and Jermiin et al. (2004) have considered the effect of ignoring changing base composition across the tree. Lockhart et al. (1998, 2000) found instances of covarion evolution where the sites that are free to vary are different in distinct lineages, and Grieben et al. (2010) showed the potential for phylogenetic error in these scenarios. Kolaczkowski and Thornton (2004) looked at the effect on phylogenetic accuracy when different partitions of the data have different branch lengths (heterotachy).

To exhibit that GTR could be causing model misspecification, we proceed as follows. Consider a single molecular sequence where each site evolves under the same, albeit heterogeneous, Markov process (we assume that sites evolve independently and under identical conditions). For time $t = 0$ to $t_1$, the sequence evolves under a time-homogeneous Markov process with substitutions governed by a rate matrix $Q_1$ whose $ij$th entry is the rate at which state $i$ changes to state $j$, so that the corresponding probability substitution matrix is given by $M_1 = e^{Q_1t_1}$. Then a disruption occurs, and over time $t_2$, the sequence again evolves under a time-homogeneous Markov process but with a different rate matrix $Q_2$ governing the substitution rates, so that the corresponding substitution matrix for this time period is $M_2 = e^{Q_2t_2}$. This scenario is illustrated in Figure 1.

As a consequence of the Markov assumption for the overall process, the substitution matrix that describes the probability of substitutions between time $t = 0$ and $t = t_1 + t_2$ can be expressed as the matrix product $\hat{M} = M_1M_2$. So far, so good, but then two questions naturally arise:

1. Is there a single rate matrix $\hat{Q}$ that can be used to describe the process from $t = 0$ to $t_1 + t_2$ as a homogeneous Markov process with $e^{\hat{Q}(t_1+t_2)} = \hat{M}?$

2. If $Q_1$ and $Q_2$ are in a particular model class and (1) is true, will $\hat{Q}$ necessarily belong to the same class?

The answer to the first question is “usually.” It is possible that, if $Q_1$ and $Q_2$ are very different, then $\hat{Q}$ may not be stochastic, that is, $\hat{Q}$ may have some negative or even complex off-diagonal entries. This possibility was noted by Baake and Haeseler (1999) and, in a slightly different context, Davies (2010) provides a useful review paper together with some illuminating examples. For the second question, Sumner et al. (2012) show that the answer is “sometimes,” depending on the model considered, and give conditions that ensure a “yes” answer. If the definition of a model class is extended to substitution matrices (some of which may not occur as the exponential of a rate matrix), the same conditions still ensure a “yes” answer to the modified question:

2’. If $M_1$ and $M_2$ are in a particular model class, will $\hat{M}$ necessarily belong to the same class?
Critical to our present discussion, Sumner et al. (2012) show that for GTR the answer to either version of the second question is a definitive “no.” Why does this lack of closure under matrix multiplication for GTR matter for phylogenetics? In almost all standard phylogenetic studies, a model is used that assumes that a homogeneous Markov process generated the molecular data. This is implemented by taking a fixed rate matrix to apply at all times and on all lineages of the evolutionary tree and helps to maximize statistical power by reducing the number of free parameters present, thus keeping the corresponding estimation variance to a minimum. However, the thoughtful phylogeneticist does not necessarily believe that the truth of the matter is that the random process remained fixed through time and across the lineages of the evolutionary tree. Rather, biological reality is likely to consist of some form of time-dependent and lineage-specific evolution governed by varying substitution rates. Under the assumption of a Markov process, this corresponds mathematically to the rate matrix changing across the evolutionary tree. For example, it has long been known that base composition is not fixed throughout evolution (Foster et al. 1997; Lockhart et al. 1992; Chang and Campbell 2000; Tarrio et al. 2001; Phillips et al. 2004) and that there is evidence that the Markov process varies across the evolutionary tree (Herbeck et al. 2005; Ota and Penny 2003).

So, it is apparent that the thoughtful phylogeneticist already thinks of her model as a statistical pragmatist’s average of the true heterogeneous process. In this light, multiplicatively closed models are then simply those where it is possible to assume a homogeneous model and “average out” these effects in a mathematically consistent way. Or, to make the point the other way around, the thoughtful phylogeneticist may be pleased to hear that for certain models it is possible to interpret her homogeneous model as providing an exact fit to a true heterogeneous process.

Before discussing how the nonclosure of GTR may affect phylogenetic estimation, we think it is worth reflecting on how GTR made its way to the top of the ModelTest hierarchy. Reversibility of a continuous-time Markov chain \( X(t) \) on some time interval \( t \in [0, T] \) arises by considering the “time-reversed process” \( Y(t) = X(T - t) \) and demanding that the probability of observing \( X(T) = j \) given \( X(0) = i \) is identical to the probability of observing \( Y(T) = i \) given \( Y(0) = j \). In this case, it is easy to show that the rate matrix for the time-reversed process \( Y(t) \) is exactly the same as that for the original process, and the process is said to be “time reversible” or simply “reversible.” Importantly, time reversibility has long been held out as important for phylogenetic analysis because implementation of Felsenstein’s “pruning” algorithm for calculating likelihoods relies on the so-called “pulley principle” and the process being reversible (Felsenstein 1981). However, recent authors have developed efficient likelihood algorithms for nonreversible processes (Boussau and Gouy 2006; Oscamou et al. 2008; Sumner and Charleston 2010), and, besides, we argue that algorithmic convenience is not the only criterion that the thoughtful phylogeneticist may wish to take into account when choosing appropriate models.

In case the concerns outlined above are seen to be somewhat philosophical, we now show that lack of multiplicative closure of a model can result in systematic biases for phylogenetic inference. If we use a model where the true heterogeneous process cannot be exactly represented homogeneously under the same model, then it is possible that there will be some error.
in estimation of substitution rates and hence species divergence times. Recalling the identifiability results of Chang (1996), any such error will lead to a misestimation of the probability distribution of site sequence patterns.

A computer simulation was conducted to test how severe this misestimation could be. As per the situation described above and represented in Figure 1, a simple situation was considered where a single sequence evolves under a homogenous GTR model before a disruption point occurs and the substitution process is allowed to change. GTR rate matrices \( Q_1 \) and \( Q_2 \) apply over times \( t_1 \) and \( t_2 \), respectively, giving the substitution matrices \( M_1 = e^{Q_1 t_1} \) and \( M_2 = e^{Q_2 t_2} \), and \( \hat{M} = M_1 M_2 \) represents the transition matrix for the entire edge. We then sought to find a GTR rate matrix \( \hat{Q} \) such that the “distance” from \( \hat{M} = e^{\hat{Q}(t_1+t_2)} \) to \( \hat{M} \) was minimized. For this purpose, we used the distance

\[
d(\hat{M}, \hat{M}) = \sqrt{\sum_{i \neq j} (\hat{M}_{ij} - \hat{M}_{ij})^2}
\]

(where the constraint \( i \neq j \) in the summation is intentional—reflecting that the diagonal entries of Markov matrices are determined by the off-diagonal entries). To measure how “heterogeneous” the process was in each case, we compared \( Q_1 \) and \( Q_2 \) using the same distance except that each rate matrix was normalized so the sum of its diagonal entries was \(-4\). To get an idea of the magnitude of this measure, we used it to compare the rate matrices that arose from a maximum-likelihood fit of GTR for each pair chosen from the data sets described in Table 2. From this, we found that values in the range 0.25–0.95 are “biological reasonable.”

In each iteration of the simulation, we set \( Q_1 \) to have all rates equal and then we selected \( Q_2 \) from the GTR model class by sampling substitution rates from a uniform distribution in such a way as to ensure a reasonably consistent spread between almost homogeneous and highly heterogeneous. All calculations were performed in the statistical computing package R (R Development Core Team 2006) and its in-built optimization routine “optim” (with the maximum number of iterations set to 4000 and relative tolerance set to \( 10^{-12} \)).

To assist in comparison, we repeated the above simulation for the known closed model F81 (for details, see Sumner et al. 2012). F81 was chosen as, despite it being guaranteed that \( \hat{Q} \) will also belong to the F81 model class (and hence our routine should find \( \hat{Q} \) such that \( e^\hat{Q} = e^{\hat{Q}} \)), it is also not the case that \( \hat{Q} \) is simply given by the weighted average of \( Q_1 \) and \( Q_2 \) (which should be compared with other closed models, such as JC69 and K80, where \( \hat{Q} \) is given by the weighted average). For this reason, F81 was deemed as a fair counterpoint to the GTR model.

For all simulations, \( t_1 \) and \( t_2 \) were chosen to be 0.5, and to ensure that the expected number of substitutions in time \( t_1 + t_2 \) was roughly in the range of 1–1.5%, we normalized both \( Q_1 \) and \( Q_2 \) by a factor that was fixed throughout. We then recorded the maximum of the percentage errors between the substitution probabilities from the true substitution matrix \( e^\hat{Q} \) to the best fit \( e^\hat{Q} \).

The results in Figure 2 show that for the F81 model, the homogeneous approximation correctly estimates the substitution probabilities to within \( 10^{-7}\% \), suggesting that the best-fit rate matrix \( \hat{Q} \) is providing the exact solution. It should be noted that the error becomes no worse as the process becomes more heterogeneous (i.e., as the distance between \( Q_1 \) and \( Q_2 \) increases), and we conclude that the errors are simply due to the optimization routine terminating.

Referring to Figure 3, we see that under the same scenario, GTR exhibited up to a 60% error in the fitted substitution probabilities, and it is also clear that the errors increase as the process is made more heterogeneous. To alleviate any concern that the poor performance of GTR in comparison with F81 was due to the former having more parameters than the latter, we significantly increased the maximum number of iterations allowed by the optimization routine and found no improvement in GTR’s performance. Recalling question (1) above, it is also worth noting that in every case the rate matrix \( \hat{Q} \) returned by the routine was stochastic.

The difference between the two cases could hardly be starker: for F81 (a closed model), the fit of a homogeneous model to a truly heterogeneous process is exact, whereas, for GTR, attempting to fit a

---

**Table 2.** Log-likelihoods of various phylogenetic models on sample data sets

<table>
<thead>
<tr>
<th>Model</th>
<th>Human</th>
<th>Acorus</th>
<th>Cormorants</th>
<th>Yeast</th>
<th>Fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>F+K (5)</td>
<td>–1557.76</td>
<td>–451,396.98</td>
<td>–7014.46</td>
<td>–710,865.35</td>
<td>–14,534.49</td>
</tr>
<tr>
<td>HKY (4)</td>
<td>–0.72</td>
<td>355.55</td>
<td>14.38</td>
<td>365.03</td>
<td>8.50</td>
</tr>
<tr>
<td>SYM (5)</td>
<td>–0.45</td>
<td>–1990.37</td>
<td>–3.11</td>
<td>–3695.09</td>
<td>5.70</td>
</tr>
<tr>
<td>GTR (8)</td>
<td>4.14</td>
<td>862.60</td>
<td>40.76</td>
<td>2409.14</td>
<td>38.86</td>
</tr>
</tbody>
</table>

Notes: The number of free parameters for each model is given in parentheses (with the count reduced by one after overall scaling is taken into account). The values for the F81+K3ST model (shortened here to “F+K”) are the actual log-likelihoods, the other values are relative to this (with positive values indicating a larger likelihood). The data sets are human, with 53 taxa, 202 sites, of mitochondrial genes (Ingman et al. 2000); Acorus, with 15 taxa, 89,436 sites, from the chloroplast genomes (Goremykin et al. 2005); cormorants, with 33 taxa, 1141 sites, from a mix of mitochondrial and nuclear genes (Holland et al. 2010); yeast, with 8 taxa, 127,026 sites, from mainly nuclear genes (Rokas et al. 2003); fish, with 11 taxa, 2178 sites, from nuclear genes (Zakon et al. 2006). Note that differences in log-likelihood greater than \( k \), where \( k \) is the difference in the number of parameters, would affect the ordering of best estimates under the Akaike information criterion.
Error in the estimate of F81 substitution probabilities for the homogeneous representation of a truly heterogeneous situation. Results are from 1000 runs, and the reasonable biological range for the x-axis is 0.25–0.95.

Error in the estimate of GTR substitution probabilities for the homogeneous representation of a truly heterogeneous situation. Results are from 1000 runs, and the reasonable biological range for the x-axis is 0.25–0.95.

As any larger model (e.g., GTR) will necessarily have more parameters and hence greater estimation variance (with the same being true for any other closed submodel of GTR). Our central proposition about GTR is not that it will fail to fit as well as its closed submodels (e.g., JC, K80, F81, K3ST)—by definition it will fit at least as well—but rather that it does not have a sensible interpretation in the context of heterogeneous sequence evolution.

Having decided that lack of closure is potentially a problem for phylogenetic models, the obvious question that arises is what are the closed models? These “good” models are not yet fully known, and Sumner et al. (2012) discuss that in general this is quite a difficult and subtle mathematical problem that crosses the boundaries of Markov chain and Lie group theory. However, they do give a method that shows how to generate a complete list of models that have a certain invariance properties under permutations of nucleotide states. In particular, they give a complete hierarchy of closed models with four-way nucleotide symmetry, that is, models that do not prefer any particular groupings of nucleotides, or to put it another way, models where for any relabeling (permutation) of the states it is possible to find a relabeling (permutation) of the substitution rate parameters such that the model is unchanged. The hierarchy they present contains GMM, F81, K3ST, and JC but includes a newly discovered six-parameter closed model dubbed “F81+K3ST” as it combines features of both these models. The rate matrices of the F81+K3ST model are most easily understood by simply summing a rate matrix from the F81 class with one from K3ST, but it should be noted that this representation is not satisfactory in general as it is technically an overparameterization (for details, see Sumner et al. 2012).

Work on cataloguing the “good” models for other symmetries is underway, with the method given in Sumner et al. (2012) being applied to generate the list of closed models that have the transition/transversion substitution symmetry. For example, K80 is an example of a model with this symmetry that is known to be closed, whereas HKY has the same symmetry but is not closed.

As a final example, we compared the performance of the closed model F81+K3ST with GTR, SYM (the GTR model with uniform base distribution), and HKY. These initial comparisons of the F81+K3ST model with models that are not multiplicatively closed show that it tends to give sometimes better and sometimes worse likelihoods. For five different data sets, a tree topology was generated using neighbor joining and then, fixing this tree topology, we performed likelihood estimation under each model, fitting substitution rates and edge lengths. The likelihood values for the various models and data sets are presented in Table 2. With the caveat that it is not necessarily statistically meaningful to compare likelihood values for these different models (especially given that the models are not nested), it can be seen from Table 2 that GTR consistently outperforms the other models, and the differences in log-likelihoods...
cannot be solely explained by its increased number of free parameters (via the Akaike information criterion, e.g.).

For the models with comparable number of parameters, we found that sometimes the best fit comes from our closed model and sometimes from a nonclosed model. What conclusions can be drawn from this study? Our overall point is that perhaps it is not prudent to only look for models that fit the data “best” in a likelihood sense but to also look for models that can be given a satisfactory interpretation. We suggest that, particularly in cases where a heterogeneous process is suspected, priority should be given to a closed model even when a nonclosed model gives a better likelihood.

To summarize, lack of closure potentially constitutes a serious problem for the use of the GTR model in phylogenetics as it means that taking an average of heterogeneous processes—which is almost certainly the underlying biological reality in many circumstances—is impossible to do in an accurate way. Further research is required to find credible closed alternatives to GTR that offer similar ability to fit phylogenetic data. It will also be important to determine closure for cases where heterogeneous models are used explicitly, for example, in codon models (Yang 1998), which are used to test for positive selection, and, for instance, the work of Hamady et al. (2006) that uses detectable shifts in inferred rate matrices to infer if genes have been horizontally transferred.

**Acknowledgments**

We thank the reviewers of our draft manuscript for comments that have greatly improved the final version of this paper. P.D.J. is an Alexander von Humboldt Fellow, J.G.Š. is an Australian Research Council (ARC) Research Fellow, and B.R.H. is an ARC Future Fellow.

**References**


