

## Opinion piece

# An examination of phylogenetic models of substitution rate variation among lineages

**Molecular evolutionary rates can show significant variation among lineages, complicating the task of estimating substitution rates and divergence times using phylogenetic methods. Accordingly, relaxed molecular clock models have been developed to accommodate such rate heterogeneity, but these often make the assumption of rate autocorrelation among lineages. In this paper, I examine the validity of this assumption.**

**Keywords:** rate autocorrelation; relaxed clocks; divergence dating; mutation rate; substitution rate; life history

## 1. INTRODUCTION

Rates of molecular evolution can vary substantially among sites, genes and lineages. In the past two decades, phylogenetic methods have been modified to take these forms of rate heterogeneity into account. For example, a number of 'relaxed-clock' models have been developed, which allow substitution rates to vary among lineages in a phylogenetic tree, without the need to assign a separate rate parameter for each branch (for a general overview, see [Rutschmann 2006](#)). These models enable the estimation of divergence times and lineage-specific substitution rates from sequence data that do not conform to a strict molecular clock. To make the estimation procedure tractable, relaxed-clock models place limitations on how rates are able to vary throughout the tree. Many of the widely used models assume that the substitution rate is indirectly heritable because it is correlated with a variety of inherited characteristics, including those associated with cellular environment, physiology and life history. Such patterns are then assumed to lead to some degree of autocorrelation between molecular rates in adjacent branches of the tree.

## 2. AUTOCORRELATED RELAXED-CLOCK MODELS

In practice, the assumption of rate autocorrelation is applied in one of several ways. In autocorrelated relaxed-clock models, the various biological factors are encapsulated in a single function describing the behaviour of rates throughout the tree. Some relaxed-

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsbl.2008.0729> or via <http://rsbl.royalsocietypublishing.org>.

One contribution to a Special Feature on 'Whole organism perspectives on understanding molecular evolution'.

clock methods employ an algorithm to minimize the rate changes between adjacent branches ([Sanderson 1997, 2002](#)), while others implement an explicit model of rate variation in which substitution rates can change or 'evolve' along branches ([Aris-Brosou & Yang 2002](#); e.g. [Huelsenbeck et al. 2000](#); [Kishino et al. 2001](#); [Lepage et al. 2006](#); [Rannala & Yang 2007](#)). After reviewing the various relaxed-clock models in detail, [Lepage et al. \(2006, 2007\)](#) proposed that the Cox–Ingersoll–Ross process possesses a number of desirable statistical properties that make it suitable for describing rate evolution. In this model, the mean rate at time  $t$ ,  $R(t)$ , is equal to

$$E[R(t)] = R(0)e^{-\theta t} + \mu(1 - e^{-\theta t}),$$

where  $\mu$  is the stationary mean of the rate and  $\theta$  determines the speed of the decay in rate autocorrelation.

Studies of simulated and real data have demonstrated that estimates of substitution rates and divergence times are sensitive to the choice of relaxed-clock model ([Ho et al. 2005](#); [Drummond et al. 2006](#); [Lepage et al. 2007](#)), highlighting the need for careful model selection.

## 3. BIOLOGICAL MOTIVATION

The biological motivation behind autocorrelated relaxed clocks can be summarized in the form of two key assumptions. The first assumption is that mutation rates are influenced by life-history characteristics such as generation time, metabolic rate and DNA repair efficiency ([Gillespie 1991](#); [Baer et al. 2007](#)). For example, herbaceous plants generally have shorter generation times than woody plants, and so exhibit higher rates of molecular evolution ([Smith & Donoghue 2008](#)).

The second assumption is that rates of mutation and substitution are correlated. Unless evolution is proceeding in an effectively neutral manner, substitution rates are somewhat removed from mutation rates. It is possible, however, that closely related species experience similar selection intensities, with comparable fitness distributions for mutations. In some empirical studies of rate variation among lineages, the two steps are not separated, and substitution rates are taken as a proxy for mutation rates. This is primarily due to the difficulty in obtaining reliable estimates of mutation rates. In any case, the substitution rate in each lineage depends on the interplay of mutation, selection and drift.

Among mammals, substitution rates have been found to be correlated with body size ([Lanfear et al. 2007](#)) or metabolic rate ([Gillooly et al. 2005](#)), synonymous rates with generation time ([Nikolaev et al. 2007](#)) and maximum lifespan ([Welch et al. 2008](#)), and non-synonymous rates with population size ([Nikolaev et al. 2007](#)) and several other traits ([Welch et al. 2008](#)). It is not clear, however, whether such patterns extend to other taxonomic groups; [Lanfear et al. \(2007\)](#) found no evidence of a metabolic rate effect on substitution rates across a variety of metazoan taxa. Investigations of the correlations between rates and biological traits in plants have yielded mixed results (e.g. [Barracough & Savolainen 2001](#); [Davies et al. 2004](#); [Smith & Donoghue 2008](#)).

Table 1. Summary of all 46 studies that have used autocorrelated relaxed clocks and have been published in Royal Society journals

journal		method <sup>a</sup>		data		regions <sup>a</sup>	
<i>Biol. Lett.</i>	5	penalized likelihood	27	DNA	43	proteins	40
<i>Phil. Trans. R. Soc. B</i>	30	non-parametric rate smoothing	20	amino acids	3	RNA	20
<i>Proc. R. Soc. B</i>	11	Bayesian relaxed clock	10			non-coding	15

<sup>a</sup>Note that the totals exceed 46 owing to the use of multiple methods and datasets in some studies.

#### 4. RATE AUTOCORRELATION IN PRACTICE

The biological assumptions underlying autocorrelated relaxed clocks warrant closer examination. The first assumption, that mutation rates are closely linked to heritable traits, receives support from studies of mammalian data. Nevertheless, even these trends differ between mammalian mitochondrial and nuclear genomes (Welch *et al.* 2008). Studies of other taxa have indicated that the correlations observed in mammals cannot be readily extended to other metazoans (Thomas *et al.* 2006; Lanfear *et al.* 2007).

Another pertinent question, related to the first biological assumption, concerns the taxonomic scale of the sequence data that are being analysed with autocorrelated relaxed-clock models. In a study of the cytochrome *b* gene in mammals, Nabholz *et al.* (2008) found that family-level categorization explained the greatest amount of rate variation. Overall, one would predict the highest degree of autocorrelation to be observed at intermediate levels of the taxonomic hierarchy. At one extreme, we would expect a very high degree of underlying rate autocorrelation within a species, such that any rate variation among lineages would be primarily due to stochastic, uninherited factors (Drummond *et al.* 2006); indeed, many population genetic and coalescent-based approaches assume a strict molecular clock.

At the other end of the continuum, autocorrelation in life-history traits (or any other factor that might be strongly correlated with mutation/substitution rates) would inevitably break down at higher taxonomic levels (Gittleman & Kot 1990; Drummond *et al.* 2006). The magnitude of the differences among lineages would be amplified if there is very incomplete taxon sampling, and the degree of autocorrelation would decrease as taxon sampling becomes more sparse. In cases where a dataset consists of distantly related taxa, there is little reason to expect any appreciable autocorrelation among the rates on different lineages. Consequently, it would be difficult to defend the validity of making *a priori* assumptions about the manner in which the rates vary among lineages.

Autocorrelated rate methods have been used to analyse sequences at various taxonomic scales, ranging from viral sequences obtained from a single host, to sequences acquired from representatives of different kingdoms of life. To investigate the trends in the application of autocorrelated relaxed clocks, a survey was conducted of all 46 studies that used such methods and were published in Royal Society journals prior to November 2008 (table 1).

The sequence data examined in these studies spanned a broad range of taxonomic levels (figure 1).

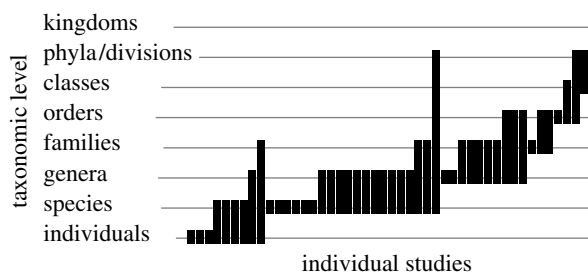


Figure 1. Plot of the approximate taxonomic levels spanned by 48 datasets that have been analysed using autocorrelated relaxed-clock models. Details of the individual studies are given in table S1 in the electronic supplementary material.

Five studies analysed datasets in which the majority of nodes in the tree represented ordinal divergences or higher. At the other extreme, nine studies involved the analyses of datasets that included large numbers of sequences from conspecific individuals, with three conducted entirely at the population level.

For the methods of analysis to be applicable to all of these datasets, they would need to be sufficiently flexible such that they could accommodate widely varying levels of rate change and autocorrelation. For small, sparsely sampled datasets, it is doubtful whether there should be any expectation of rate autocorrelation at all.

The second assumption behind autocorrelated relaxed-clock models is that mutation and substitution rates are strongly correlated. This is reasonable for sequences that are evolving neutrally. In analyses of sequences under selection, however, such an assumption is far more questionable. This relates particularly to non-mammalian mitochondrial sequence data, of which the evolutionary history appears to have been driven substantially by adaptive evolution (Bazin *et al.* 2006). If rates of adaptive substitution are not tied to inherited factors, then the presence of such substitutions can seriously weaken the link between life-history traits and substitution rates. As mentioned above, however, closely related species could experience similar selection intensities, as implied under covarion models of sequence evolution (e.g. Tuffley & Steel 1998). The extent to which such processes could lead to rate autocorrelation among lineages is not known.

In a comprehensive study of mammals, no correlation was found between non-synonymous mitochondrial rates and life-history traits (Welch *et al.* 2008). Indeed, this suggests that autocorrelated relaxed-clock models might be inappropriate for analyses of amino acid sequences. Thus, perhaps it would be desirable to employ separate autocorrelated and uncorrelated models of among-lineage rate



- Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc. Natl Acad. Sci. USA* **104**, 20 443–20 448. (doi:10.1073/pnas.0705658104)
- Rannala, B. & Yang, Z. 2007 Inferring speciation times under an episodic molecular clock. *Syst. Biol.* **56**, 453–466. (doi:10.1080/10635150701420643)
- Rutschmann, F. 2006 Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. *Divers. Distrib.* **12**, 35–48. (doi:10.1111/j.1366-9516.2006.00210.x)
- Sanderson, M. J. 1997 A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**, 1218–1231.
- Sanderson, M. J. 2002 Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109.
- Seo, T. K., Kishino, H. & Thorne, J. L. 2004 Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Mol. Biol. Evol.* **21**, 1201–1213. (doi:10.1093/molbev/msh088)
- Smith, S. A. & Donoghue, M. J. 2008 Rates of molecular evolution are linked to life history in flowering plants. *Science* **322**, 86–89. (doi:10.1126/science.1163197)
- Thomas, J. A., Welch, J. J., Woolfit, M. & Bromham, L. 2006 There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. *Proc. Natl Acad. Sci. USA* **103**, 7366–7371. (doi:10.1073/pnas.0510251103)
- Tuffley, C. & Steel, M. 1998 Modeling the covarion hypothesis of sequence evolution. *Math. Biosci.* **147**, 63–91. (doi:10.1016/S0025-5564(97)00081-3)
- Welch, J. J., Bininda-Emonds, O. R. & Bromham, L. 2008 Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol. Biol.* **8**, 53.