# Local and relaxed clocks, the best of both worlds

**Mathieu Fourment** Corresp., [1] , **Aaron E Darling** [1]

[1] ithree institute, University of Technology Sydney, Sydney, Australia

Corresponding Author: Mathieu Fourment
Email address: mathieu.fourment@uts.edu.au

Time-resolved phylogenetic methods use information about the time of sample collection to estimate the rate of evolution. Originally, the models used to estimate evolutionary rates were quite simple, assuming that all lineages evolve at the same rate, an assumption commonly known as the molecular clock. Richer and more complex models have since been introduced to capture the phenomenon of substitution rate variation among lineages. Two well known model extensions are the local clock, wherein all lineages in a clade share a common substitution rate, and the uncorrelated relaxed clock, wherein the substitution rate on each lineage is independent from other lineages while being constrained to fit some parametric distribution. We introduce a further model extension, called the flexible local clock (FLC), which provides a flexible framework to combine relaxed clock models with local clock models. We evaluate the flexible local clock on simulated and real datasets and show that it provides substantially improved fit to an influenza dataset. An implementation of the model is available for download from https://www.github.com/4ment/flc.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26744v1 | CC BY 4.0 Open Access | rec: 21 Mar 2018, publ: 21 Mar 2018

# Local and relaxed clocks, the best of both worlds

**Mathieu Fourment**[1] **and Aaron E. Darling**[1]

[1]**The ithree institute, University of Technology Sydney, Ultimo, NSW, 2007, Australia**

Corresponding author:
Mathieu Fourment[1]

Email address: mathieu.fourment@uts.edu.au

## ABSTRACT

Time-resolved phylogenetic methods use information about the time of sample collection to estimate the rate of evolution. Originally, the models used to estimate evolutionary rates were quite simple, assuming that all lineages evolve at the same rate, an assumption commonly known as the molecular clock. Richer and more complex models have since been introduced to capture the phenomenon of substitution rate variation among lineages. Two well known model extensions are the local clock, wherein all lineages in a clade share a common substitution rate, and the uncorrelated relaxed clock, wherein the substitution rate on each lineage is independent from other lineages while being constrained to fit some parametric distribution. We introduce a further model extension, called the flexible local clock (FLC), which provides a flexible framework to combine relaxed clock models with local clock models. We evaluate the flexible local clock on simulated and real datasets and show that it provides substantially improved fit to an influenza dataset. An implementation of the model is available for download from `https://www.github.com/4ment/flc`.

## INTRODUCTION

Phylogenetic methods provide a powerful framework for reconstructing the evolutionary history of viruses, bacteria, and other organisms. Correctly estimating the rate at which mutations accumulate in a lineage is essential for phylogenetic analysis, as the accuracy of inferred rates can heavily impact other aspects of the analysis. Classic approaches to infer the substitution rate of a group of organisms rely on the existence of a so-called ”molecular clock”. The molecular clock hypothesis dictates that mutations accumulate at an approximately steady rate over time, implying that the genetic distance between two organisms is proportional to the time since these organisms last shared a common ancestor. The molecular clock hypothesis was first proposed almost 50 years ago by Emile Zuckerkandl and Linus Pauling (Zuckerkandl and Pauling, 1965) who suggested that the substitution rate was effectively constant over time. This very restricted model of evolution has been implemented using a ”strict clock” model in phylogenetic inference software, but the rates of evolution in many organisms appears to change over time and the model can not capture this phenomenon.

In recent years, richer models have been developed to capture the complexity of the evolutionary process. Sanderson (2002) and Thorne et al. (1998) proposed to model rate heterogeneity among lineages using auto-correlated clock models using penalized likelihood and Bayesian inference respectively. In these parameter rich models, the rate of each lineage is assumed to be correlated to that of the parent lineage. The auto-correlation assumption could be justified by considering that the substitution rate is influenced by heritable mechanisms such as metabolic rate or generation time. However there is no guarantee that rates evolve in an auto-correlated manner, especially when the timescale under study is relatively small Drummond et al. (2006). An alternative approach is to assume that substitution rates on adjacent branches are independent draws from an underlying parametric distribution. Drummond et al. (2006) chose to forgo the hierarchical Bayesian framework and opted for a likelihood approach that requires the rates to fit a discretized distribution. The log-normal and exponential distributions are commonly used as they are available in the widely used BEAST package (Drummond and Rambaut, 2007; Bouckaert et al., 2014). The auto-correlated and uncorrelated clock models are referred to as *relaxed*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26744v1 | CC BY 4.0 Open Access | rec: 21 Mar 2018, publ: 21 Mar 2018

*clock models* due their ability to relax the constant rate assumption.

Local clock models are an alternative to relaxed clocks, where the model assumes that the substitution rate is constant within a clade but can differ between clades (Yoder and Yang, 2000; Yang and Yoder, 2003). The number and locations of these local clocks can be inferred from the data using the random local clock model (Drummond and Suchard, 2010) or local clocks can be assigned by the user based on prior information.

In this manuscript we introduce a hybrid model that integrates features of both the local and the relaxed clock models. In the model each local clock can be specified either as a strict clock (as in the original formulation of the local clock model) or as a relaxed clock. We call this model the flexible local clock (FLC) model. We evaluate the FLC model using a newly implemented module for the BEAST2 package, which uses Markov chain Monte Carlo to carry out inference of model parameters(Bouckaert et al., 2014). We reanalyzed an influenza virus (Drummond and Suchard, 2010) and a HIV (Wertheim et al., 2012) data set to evaluate the utility of the FLC model and compared its fit to the data to that given by other models.

## METHODS

Phylogenetic packages such as BEAST provide several options to model lineage-specific rate variation, known as heterotachy, without overfitting the model. One of the first ingredients of the FLC model is the uncorrelated relaxed clock model Drummond et al. (2006), arguably the most popular lineage-specific rate model. The uncorrelated relaxed clock model uses a single discretized parametric distribution to model rate heterogeneity. In the original formulation of the model, a parametric distribution, usually lognormal, is discretized into a fixed number of components, with the number of these components equal to the number of branches $b$ in the tree. In its simplest form, the model assumes a one-to-one relationship between a rate at a branch and one of the components. For a lognormal distribution, this approach only requires estimating two parameters (i.e. mean and standard deviation) instead of $2N-2$ parameters if a hierarchical model was used, where $N$ is the number of sequences.

The other ingredient of the FLC model is the local clock which was first proposed by Yoder and Yang (2000). This model allows lineages within a region of the tree to evolve at exactly the same rate. We define a local clock on a phylogeny as a monophyletic group where the substitution rate of every lineage is equal. As in Drummond and Suchard (2010), we assume the existence of another clock (e.g. a 'global' clock) for lineages that are not assigned a local clock.

Herein, we propose to relax the constraint that lineages within a local clock evolve at exactly the same rate by replacing this implicit strict clock by a relaxed clock.

We applied the FLC model to two data sets of heterochronous viral nucleotide sequences. The first data set comprises an alignment of 69 human influenza A/H3N2 virus haemagglutinin (HA) sequences (987 nt in length) isolated between 1981 and 1998. The evolutionary rates and time to the most recent ancestors (tMRCAs) of this data set was previously investigated using a random local clock method (Drummond and Suchard, 2010) with a Bayesian Markov chain Monte Carlo (MCMC) approach implemented in BEAST (Drummond and Rambaut, 2007). We reanalysed the data using BEAST with either the FLC model, uncorrelated lognormal relaxed clock (UCLN), local clock (LC), and a random local clock model (RLC). As in the original study, our analyses use the HKY+$\Gamma_4$ substitution model that incorporates gamma-distributed rate variation among sites (4 rate classes). The FLC and LC models require manual assignment of each lineage to a local clock with the appropriate constraints on the phylogeny. Drummond and Rambaut (2007) noticed that the substitution rate of the lineages comprising viruses sampled after 1990 appeared higher than the pre-1990 lineages. We therefore assigned sequences sampled after 1990 to a local clock for both LC and FLC models. For each local-based model, we conducted two separate analyses in which the branch subtending the clade containing the late viruses (1990-onward) were assigned either to a local clock or the ancestral rate. We specified a diffuse prior on the substitution rates of the LC and FLC models using an exponential distribution with a mean of 0.003. For the log-normal distribution of the relaxed clock we used an exponential prior distribution ($\lambda = 1/0.003$) on the mean parameter and an exponential prior distribution ($\lambda = 1/0.33$) on the standard deviation parameter. As in the study describing the RLC model we used a Poisson distribution with $\lambda = \log 2$ as a prior on the number of local clocks, thereby placing 50% prior probability on a single rate across the phylogeny. Finally, we assumed *a priori* that rate multipliers are independently gamma distributed with $\alpha = 0.5$ and $\beta = 2$ as in Drummond and Suchard (2010).

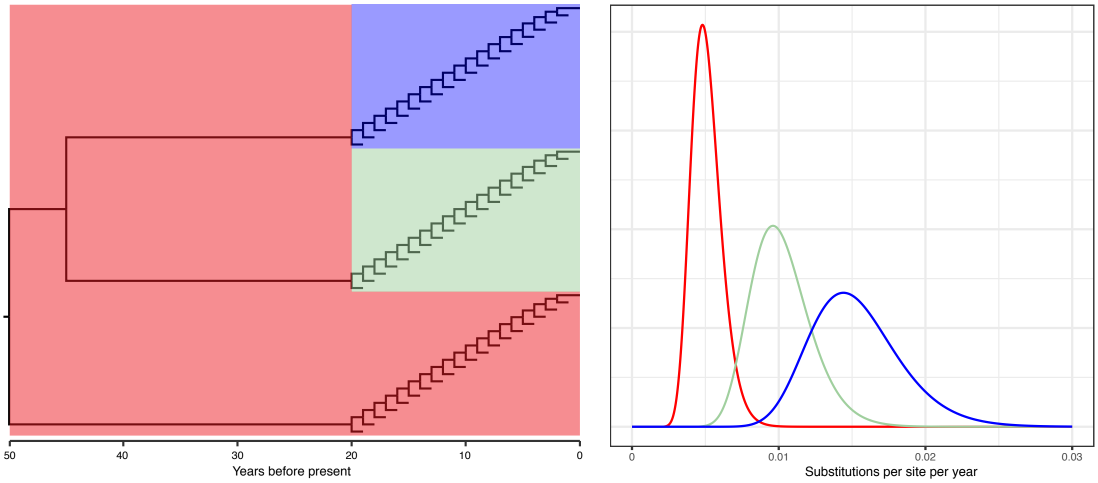

**Figure 1.** Phylogenetic tree and substitution rates used to simulate data sets.

For each data set, we calculated the marginal likelihood of the data under each model using the stepping stone algorithm to compare competing models Xie et al. (2011). We used a series of 100 power posteriors where $\beta$ values are chosen to be evenly spaced quantiles of a Beta distribution with parameters $\alpha = 0.3$ and $\beta = 1.0$. These parameters results in half of the power posteriors being evaluated for $\beta < 0.1$ for which the power posterior is changing rapidly, as suggested by Xie et al. (2011). Each MCMC was run for 10 million iterations and the first 10% of the samples were discarded as burn-in.

### Simulations

To validate the implementation of our model we simulated data sets using the FLC model. Our approach is similar to a simulation-based study (Worobey et al., 2014) that showed that the local clock model is best suited to model rate variation among influenza virus sequences sampled from three different hosts (i.e. equine, human, and birds). Worobey *et al.* assigned different local clocks to each of the monophyletic bird and human clades and simulated nucleotide alignments containing 10,000 sites. Phylogenies were estimated using either a strict, flexible local or local clock model using the BEAST package (Drummond and Rambaut, 2007). The simulations showed that only the local clock model was able to recover the true tree. In this study, we used the same topology and divergence times, and replaced standard local clocks with flexible local clocks.

10 replicates containing 10,000 sites were simulated using the program simultron (Fourment and Holmes, 2014) under the HKY model ($\kappa = 3$ and equal nucleotide frequencies). The standard deviation $\sigma$ of the lognormal distributions were all set to be equal to 0.2. The $\mu$ parameter of the lognormal distributions of the equine, human, and bird clades were set such as the mean of the distributions were $5 \times 10^{-3}$, $1 \times 10^{-2}$, and $1.5 \times 10^{-2}$ respectively (Figure 1). The choice of the parameters results in roughly bell shaped distributions centered on the substitution rates used in the Worobey *et al.* study. We analyzed the simulated data sets with the HKY model and the skyline coalescent tree prior under the strict, flexible local, local, relaxed, and random local clocks. The simulation script is available from `http://www.github.com/4ment/flc-data`.

## RESULTS AND DISCUSSION

We analyzed the influenza virus data set with BEAST under a variety of models including the FLC model. Since the flexible local clock can be composed of a combination of strict and relaxed clocks, we specify the type of clock between brackets. For example, we use FLC [strict&UCLN] to denote a flexible local clock with a strict clock on the early lineages (i.e. sequences before 1990) and a uncorrelated lognormal relaxed clock (UCLN) on the later lineages. For local and flexible local clocks we can specify whether the branch leading to the clade with a local clock should be included in the new clock (contains the stem). To test which models better fit to the data we calculated the marginal likelihood for each model (Table 1).

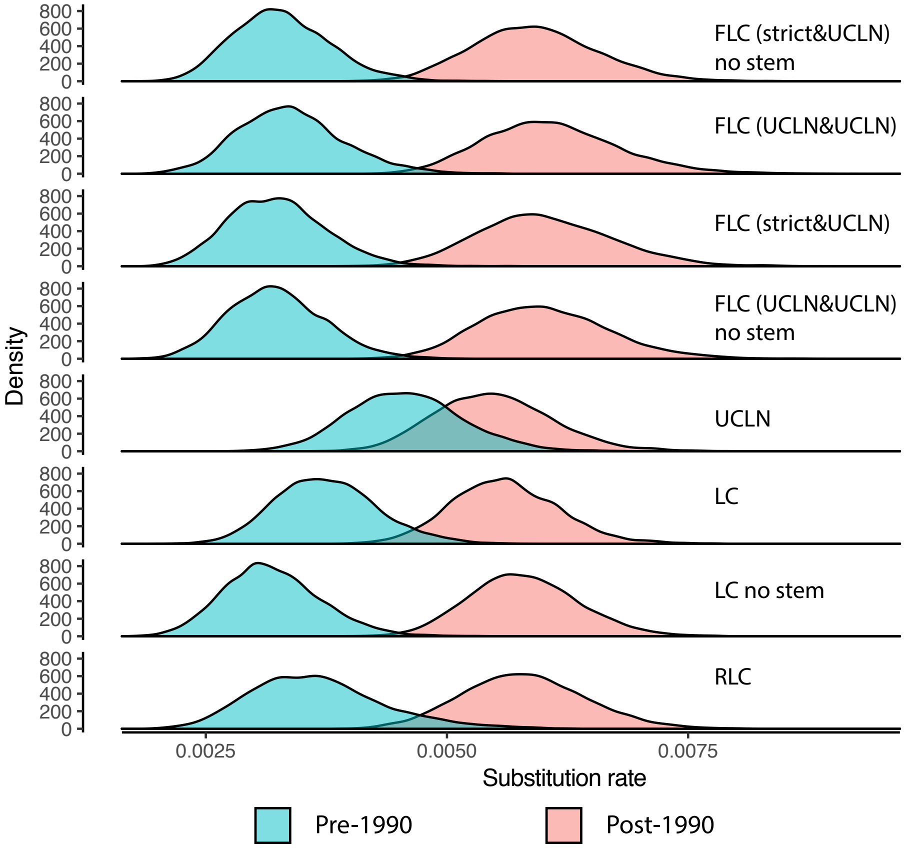

**Figure 2.** Posterior distributions of the mean substitution rate of the lineages comprising viruses sampled after 1990 and the pre-1990 lineages. UCLN: uncorrelated lognormal relaxed clock, RLC: random local clock, FLC: flexible local clock LC: local clock. For the local clock models labeled no stem, the branch subtending the post 1990 clade is not assigned to the local clock.

| Model | Marginal likelihood | Contains stem |
|---|---|---|
| FLC [strict&UCLN] | -4381.72 | no |
| FLC [UCLN&UCLN] | -4382.07 | yes |
| FLC [strict&UCLN] | -4382.92 | yes |
| FLC [UCLN&UCLN] | -4383.28 | no |
| UCLN | -4385.04 | NA |
| LC | -4386.81 | yes |
| LC | -4387.69 | no |
| RLC | -4415.36 | NA |

**Table 1.** Marginal likelihoods calculated using the stepping stone algorithm. UCLN: uncorrelated lognormal relaxed clock, RLC: random local clock, FLC: flexible local clock LC: local clock. The ”Contains stem” column specifies whether the branch subtending the post-1990 clade is assigned to the local clock.

As in the original study (Drummond and Suchard, 2010), every model shows a substitution rate increase in sequences sampled after the 1990 (Figure 2).

The marginal likelihood estimates (Table 1) suggest that the best models are the FLC models, followed by the UCLN, LC, and RLC models. The inclusion of the stem in the FLC and LC models appears to a minor effect on the model fit depending on the model, but the marginal likelihood estimates are subject to Monte Carlo error and caution should be exercised in order to avoid overinterpreting small differences. The 95% highest posterior density (HPD) of the standard deviation of the lognormal distribution assigned to the global clock includes zero, suggesting that there is little rate variation outside the post-1990 clade (i.e. FLC [UCLN&UCLN]). It is therefore no surprise that the marginal likelihoods of the FLC models with a strict or UCLN clock on the pre-1990 lineages are similar. Interestingly the UCLN model appears to fit better to the data than the RLC and LC clocks.

## Results on simulated data

We simulated 10 data sets under the flexible local model and estimated the phylogenies using several clock models. The comparison of the maximum clade credibility (MCC) tree to the true topology reveals that the strict and relaxed clock models could not recover the rooting of the true tree in any replicate. Interestingly the 95% HPD intervals of the root node age contained the true value in four and two of the replicates using the relaxed and strict clock models, respectively. The MCC trees of the standard local clock model recovered the true rooting and the root age was recovered in the 95% HPD in only three replicates. The MCC trees of the flexible local clock model had the same rooting as the true tree in 9 out of 10 cases and the 95% HPD of the root age contained the true value for 8 out of 10 replicates.

## Limitations and conclusions

As in the standard local model, the flexible local clock model introduced in this paper assumes that the user knows the number and the location of the rate shifts. Drummond and Suchard (2010) devised the random local clock to address this limitation using a stochastic search variable selection method to sample over random local clocks. Unfortunately that approach is not easily amenable to integration with the FLC model since the substitution rate within a clock can either be constant or heterogeneous across lineages. Although it should be possible to use reversible jump MCMC to sample the posterior distribution it is not clear how to deal with a variable number of lineages assigned to a relaxed clock. For example, the assignment of a relaxed clock with a two parameter distribution to a single branch would over-parametrize the model. An interesting direction for further research would be to develop an algorithm that automatically selects the clock type for each local clock.

The FLC model is implemented in the BEAST 2 package as a plugin and is available from `https://www.github.com/4ment/flc`.

# REFERENCES

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7:214.

Drummond, A. J. and Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biol*, 8:114.

Fourment, M. and Holmes, E. C. (2014). Novel non-parametric models to estimate evolutionary rates and divergence times from heterochronous sequence data. *BMC evolutionary biology*, 14(1):163.

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol Biol Evol*, 19(1):101–9.

Thorne, J. L., Kishino, H., and Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*, 15(12):1647–57.

Wertheim, J. O., Fourment, M., and Kosakovsky Pond, S. L. (2012). Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Mol Biol Evol*, 29(2):451–6.

Worobey, M., Han, G.-Z., and Rambaut, A. (2014). A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Syst Biol*, 60(2):150–60.

Yang, Z. and Yoder, A. D. (2003). Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst Biol*, 52(5):705–16.

Yoder, A. D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol*, 17(7):1081–90.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, pages 97–166.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26744v1 | CC BY 4.0 Open Access | rec: 21 Mar 2018, publ: 21 Mar 2018