On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference

Vivek Gowri-Shankar and Magnus Rattray

School of Computer Science, University of Manchester, Manchester M13 9PL, UK

Research article

Running head: Compositional heterogeneity across sites

Corresponding author:
Magnus Rattray
School of Computer Science
University of Manchester
Manchester, M13 9PL
Phone: +44 (0)161 275 6187
Fax: +44 (0)161 275 6202
magnus.rattray@cs.man.ac.uk

Keywords: phylogeny, RNA, G+C content, compositional heterogeneity, across-site variation, ancestral composition
Abstract

Model-based phylogenetic reconstruction methods traditionally assume homogeneity of nucleotide frequencies among sequence sites and lineages. Yet heterogeneity in base composition is a characteristic shared by most biological sequences. Compositional variation in time, reflected in the compositional biases among contemporary sequences, has already been extensively studied and its detrimental effects on phylogenetic estimates are known. However, fewer studies have focused on the effects of spatial compositional heterogeneity within genes. We show here that different sites in an alignment do not always share a unique compositional pattern and we provide examples where nucleotide frequency trends are correlated with the site-specific rate of evolution in RNA genes. Spatial compositional heterogeneity is shown to affect the estimation of evolutionary parameters. With standard phylogenetic methods, estimates of equilibrium frequencies are found to be biased towards the composition observed at fast evolving sites. Conversely, the ancestral composition estimates of some time-heterogeneous but spatially homogeneous methods are found to be biased towards frequencies observed at invariant and slow evolving sites. The latter finding challenges the result of a previous study arguing against a hyperthermophilic last universal ancestor from the low apparent G+C content of its rRNA sequences. We propose a new model to account for compositional variation across sites. A Gaussian process prior is used to allow for a smooth change in composition with evolutionary rate. The model has been implemented in the phylogenetic inference software PHASE and Bayesian methods can be used to obtain the model parameters. The results suggest that this model can accurately capture the observed trends in present-day RNA sequences.
Introduction

Contemporary phylogenetic methods are no longer limited to the descriptive study of species relatedness. Current model-based approaches are as much designed to unravel the mechanisms of sequence evolution as they are to recover the evolutionary relationship of organisms and the Markov models that describe the nucleotide substitution process have become increasingly complex (see Felsenstein (2004) for a review). In an attempt to accurately model the biological processes involved, the incorporation of among-site rate variation (ASRV) in substitution models was an important step towards more realistic evolutionary models (Yang, 1996a; Felsenstein, 2001). Indeed, the selection pressure does not act uniformly over the length of a gene and some sites change often over evolutionary time whereas others are almost invariant due to strong functional constraints (Kimura and Otha, 1974). The rate at which a mutation becomes fixed in a population varies depending on its position in the sequence. Nevertheless, current ASRV methods only account for the variation of the site-specific speed of the nucleotide replacement process while other evolutionary patterns are assumed to be shared across sites: the main parameters of the Markov process (namely equilibrium frequencies and exchangeabilities) are constant for the whole alignment. One point we wish to develop in this article is that the combined effects of mutation and selection also have an impact on nucleotide frequencies that current ASRV models cannot capture.

There are cases where evolutionary patterns are known to be different within a gene (e.g., different codon positions of a protein coding gene or loops and stems of RNA genes). An easy way to account for this heterogeneity is to partition the data beforehand according to a priori knowledge. Combined models that use different substitution models for different partitions have been studied and are in common use (Yang, 1996b; Pupko et al., 2002; Seo et al., 2005).

Nevertheless, partitioning of the data is only an option when partitions can be
identified in advance and we are more interested here in compositional heterogeneity within a partition. In the past ten years, methods have been proposed to capture spatial heterogeneity (not specifically compositional heterogeneity) when the correct partitioning scheme is unknown or when intragenic variability cannot be related directly to specific DNA segments. Latent class models (or mixture models) have been used quite often in this context. These models assume that each site is evolving according to an unobserved process chosen among a finite set of substitution models and the likelihood is computed by integrating over all the possible substitution processes at a site. For instance, Huelsenbeck and Nielsen (1999) introduced a model where the transition/transversion rate ratios can vary along a sequence according to a gamma distribution. Following Nielsen and Yang’s (1998) work, Yang et al. (2000) proposed a set of codon models with variable synonymous/nonsynonymous ratio across sites. The standard discrete gamma rate model for ASRV (Yang, 1994) and the invariant sites model (Reeves, 1992) were early examples of latent class models that allow for rate-heterogeneity across sites without prior classification.

Closer to the main topic of this paper, some methods have also been devised specifically to account for variation in composition across sites. Most substitution models ignore this aspect of sequence evolution and assume that the equilibrium frequencies across sites are constant. Dimmic et al. (2000) incorporated site-specific selection effects using different amino-acid fitness functions. More recently, Pagel and Meade (2004) introduced a “pattern-heterogeneity” mixture model to account for the heterogeneity both in average evolutionary rate and exchangeability parameters, and their model can easily be extended to describe the spatial variation of frequency parameters. Both models are also latent class methods. Models using a specific frequency vector at each site have also been designed. Lartillot and Philippe (2004) constructed a model that allows for a variable number of frequency profiles to fit the variation of the amino-acid equilibrium distribution and, in effect, classifies sites into distinct categories. A parameter-rich model was also attempted by Bruno
(1996) and Halpern and Bruno (1998), where a single mutation process is shared across sites but a site-specific frequency vector is used to model the variation of selection effects across sites.

In this paper we concentrate on variation in nucleotide composition across RNA genes. Secondary structure constraints on RNA molecules and differences between paired and unpaired regions are taken into account. By contrasting the composition observed at slow and fast evolving pairs in contemporary RNA helices, it is suggested that the differences between their patterns of evolution is not simply limited to variation of the average substitution rate. The most stable G:C pairs are more common in slowly evolving parts of RNA stems and nucleotide frequencies in RNA loops are also found to be quite variable across rate categories. In general it appears that base composition is correlated with the site-specific evolutionary rate. In order to account for the observed compositional trends across sites in present-day sequences, we propose a new latent class method for ASRV. Our model extends the discrete gamma model (Yang, 1994) to allow for variable equilibrium frequencies in each rate category. The main model assumption, supported by empirical evidence, is that these frequencies change in a smooth way across rate categories. We use a Gaussian process to control the smoothness and parameters are estimated from the data during a Bayesian inference process using Markov chain Monte Carlo (MCMC) techniques. This model is implemented and available in PHASE, a software package for Maximum Likelihood (ML) and Bayesian phylogenetic inference designed specifically for use with RNA genes (Jow et al., 2002; Hudelot et al., 2003).

We motivate our work by highlighting the biasing effects of compositional heterogeneity across sites on standard phylogenetic methods that do not account for it. In our previous works, we noticed that frequency estimates were biased towards the frequencies of fast evolving sites (Jow et al., 2002; Hudelot et al., 2003) and we investigated whether compositional variations across sites could explain these results. More generally, the effects of compositional heterogeneity on phylogenetic estimates
is explored in the ML and Bayesian frameworks.

Since lineage-specific base compositional bias has previously been shown to cause phylogenetic artifacts (Lockhart et al., 1994; Jermiin et al., 2004), more effort has been put into modelling compositional variation over evolutionary time rather than across sites (Yang and Roberts, 1995; Galtier and Gouy, 1998; Brooks et al., 2004; Foster, 2004). However, we find that the impact of compositional variation across sites on models proposed to relax the assumption of time-homogeneity is also worrying. We show, as an example, that under a model that wrongly assumes across-site homogeneity of the equilibrium frequencies, part of the observed variation across time will in fact be due to variations across sites, especially when the model is used conjointly with a model of rate heterogeneity. Galtier et al. (1999) suggested the low inferred G+C content of the Last Universal Common Ancestor (LUCA) of all extant life forms was not compatible with the expected G+C content of a thermophilic species but this result has been challenged by several other studies (Di Giulio, 2000, 2003; Schwartzman and Lineweaver, 2004). We provide results suggesting that the ancestral G+C compositions proposed by Galtier et al. (1999) for the two studied rRNAs genes were most likely underestimated.

**Materials and methods**

Throughout the paper, two very different methods are used to estimate frequencies at slow and fast evolving sites and to produce “frequency vs. average rate” curves. The first method is a fast ML method that estimates the parameters of a standard site-homogeneous model (but allowing for rate variation across sites) and computes rate-specific empirical Bayes posterior averages. The second method uses a site-heterogeneous model that we develop in this work. Both methods are described at the end of the materials and methods section after a description of the evolutionary models.
Substitution models

Standard Markov models of nucleotide substitution are used in this paper (Swofford et al., 1996). These models are fully defined by a matrix $Q = \{t_{ij}\}$ of transition rates between states (e.g., the four nucleotides or the twenty amino-acids). We typically assume that the process is reversible and, consequently, it can be parameterized by a set of stationary frequencies, $\Pi = \{\pi_i\}$, which is the expected composition when the process is at equilibrium, and a symmetric matrix of exchangeability parameters $R = \{\rho_{ij}\} = \{\rho_{ji}\}$. Using these parameters, the instantaneous transition rates of the Markov process are defined as:

$$
t_{ij} = \mu \pi_j \rho_{ij} \quad \forall (i, j)/i \neq j,
$$

$$
t_{ii} = -\mu \sum_{j=1, j \neq i}^{n} \pi_j \rho_{ij} \quad \forall i,
$$

where $\mu$ is a scaling factor determined by the desired average substitution rate of the process.

The most general time-reversible model for nucleotide substitution has four frequency parameters and six exchangeability parameters. Frequencies must sum up to 1.0 and we fix $\rho_{AG}$ to 1.0 because it is not possible to identify separately the six exchangeability parameters and $\mu$. For computational reasons, and because it does not affect the arguments we are developing in this paper, we will use a constrained version of this general model proposed by Tamura and Nei (1993). It has two exchangeability rates $\rho_{AG}$ and $\rho_{CT}$ for nucleotide transitions (with $\rho_{AG} = 1.0$) and only one exchangeability rate for nucleotide transversions ($\rho_{AC} = \rho_{AT} = \rho_{CG} = \rho_{CT}$).

RNA molecules fold onto themselves to produce small stretches of base-pairing (helices or stems) and single-stranded loops. Watson-Crick pairs (A:U and G:C) and the intermediate non-canonical G:U pair are most commonly found in RNA helices. Consequently, RNA sequences are associated with a secondary structure that turns out to be well conserved over evolutionary time, even though the un-
derlying nucleotide sequences are not. To preserve RNA helices, compensatory 
substitutions occur frequently to maintain complementarity (Stephan, 1996). Re-
placements within the two symmetric groups [A:U, G:U, G:C] and [U:A, U:G, C:G] 
are quite common whereas exchanges between these two groups are unlikely since 
they involve an intermediate state that is thermodynamically unstable. As far as 
phylogenetics is concerned, the compensatory substitution mechanism affects stan-
dard methods since it is traditionally assumed that sites are evolving independently 
and spatially distant substitutions are assumed to be uncorrelated. Specific sub-
titution models that account for the RNA secondary structure are implemented in 
PHASE (http://www.bioinf.man.ac.uk/resources/phase). Reviewed by Savill et al. 
(2001), these models consider pairs of nucleotides as their elementary unit and par-
tially alleviate the independence assumption. In this paper, we use the OTRNA 
model introduced by Tillier and Collins (1998), and referred to as 7D in Savill et al. 
(2001). The transition matrix is given below. This model is a 7-state model with 
the four Watson-Crick pairs (A:U/U:A and G:C/C:G) and the two U:G/G:U pairs. 
The ten remaining unstable combinations are lumped into a single mismatch state 
MM. The OTRNA model has seven frequency parameters. Exchangeabilities corre-
spanding to each other in the two groups defined above are assumed to be equal, 
et. g., $\rho_{A:U,G:U} = \rho_{U:A,U:G}$, and the model has only four different exchangeability pa-
rameters. Inside each symmetric group, one parameter is used for the two single 
transitions $\alpha_s = \rho_{A:U,G:U} = \rho_{G:U,G:C}$ and one for the double transition $\alpha_d = \rho_{A:U,G:C}$, 
set to 1.0 as a reference. The third exchangeability parameter $\beta$ is for the unlikely 
replacements between the two groups defined above while the fourth one $\gamma$ is for the
replacements to and from the mismatch state.

\[
Q = \mu \times \begin{pmatrix}
AU & GU & GC & UA & UG & CG & MM \\
AU & * & \alpha_s\pi GU & \pi GC & \beta\pi UA & \beta\pi UG & \beta\pi CG & \gamma\pi MM \\
GU & \alpha_s\pi AU & * & \alpha_s\pi GC & \beta\pi UA & \beta\pi UG & \beta\pi CG & \gamma\pi MM \\
GC & \pi AU & \alpha_s\pi GU & * & \beta\pi UA & \beta\pi UG & \beta\pi CG & \gamma\pi MM \\
UA & \beta\pi AU & \beta\pi GU & \beta\pi GC & * & \alpha_s\pi UG & \pi CG & \gamma\pi MM \\
UG & \beta\pi AU & \beta\pi GU & \beta\pi GC & \alpha_s\pi UA & * & \alpha_s\pi CG & \gamma\pi MM \\
CG & \beta\pi AU & \beta\pi GU & \beta\pi GC & \pi UA & \alpha_s\pi UG & * & \gamma\pi MM \\
MM & \gamma\pi AU & \gamma\pi GU & \gamma\pi GC & \gamma\pi UA & \gamma\pi UG & \gamma\pi CG & * 
\end{pmatrix}
\]

**Across-sites heterogeneity**

RNA genes were partitioned according to their secondary structure in order to perform combined analyses. Single-stranded and helical regions are modelled using different evolutionary models. A standard nucleotide model is used with unpaired nucleotides and a specific paired-site model is used for the helices. Since secondary structure can change over evolutionary time, a consensus structure must be used. Pairs that are not conserved for at least 50% of the species are not retained and the two nucleotides are considered as independently evolving sites in this case. In a combined analysis, substitution models can share some parameters (Yang, 1996b) but we assume here that only the phylogeny is shared. The two partitions are supposed to have evolved along a common topology and it is also assumed that the average substitution rate of each substitution model are related by a proportionality factor that is constant over the whole tree. This is in effect equivalent to Yang’s (1996b) model, which assumes that branch lengths for each partition are related by a proportionality factor.

We use the discrete gamma model to account for ASRV (Yang, 1994). Across-site rate variation is modelled by a gamma distribution, approximated for computational
reasons by a finite number of categories $k$ with equal prior probabilities $1/k$. Two parameters control the shape of this distribution: a mean and a shape parameter. The mean is the average substitution rate of the model, i.e., the expected number of substitutions over a branch of length 1.0. The gamma shape parameter ($\alpha$) determines the extent of rate variation.

**Sequence data**

We use both real and synthetic datasets. The real RNA alignments D1 and D2 are distributed with the PHASE package or available upon request. The synthetic datasets D3 and D4 were generated by PHASE and used to investigate spatial compositional heterogeneity using simulations. Later on, when mentioning D3 or D4, we actually refer to a replicate generated according the evolutionary models described below and the number of sites in the alignment will always be specified.

**D1: Mitochondrial tRNA and rRNA genes from primates**

This dataset was extracted directly from the dataset used for mammalian phylogenetic inference by Hudelot et al. (2003). The genes used are the complete set of mitochondrial tRNAs and rRNAs. For computational reasons, we kept only 13 primate species (gorilla, human, chimpanzee, pygmy chimpanzee, orangutan, gibbon, baboon, barbary ape, capuchin, loris, tarsier, ring-tailed lemur, malayan flying lemur) and three species from the grouping Laurasiatheria as an outgroup (dog, cow and rhinoceros) from the original dataset. The consensus secondary structure is a key part of this alignment since nucleotides are treated differently according to their position. This dataset is used to illustrate compositional variation across sites in real sequences and we assume that the substitution process is not excessively time-heterogeneous with this alignment.

**D2a & D2b: Nuclear LSU and SSU RNAs**

These two rRNA genes were used by Galtier et al. (1999) to infer the G+C composition of the ancestral sequence of the common ancestor to all life forms and
deduce its likely environmental temperature. The sequences were downloaded from the European Ribosomal RNA database (Wuyts et al., 2001, 2002) for the same set of 40 species. Ambiguous regions were removed and the alignments were further refined by eye. Once the biasing effects of compositional variation across sites on time-heterogeneous methods are demonstrated, this dataset, highly heterogeneous both in time and in space, is used to reevaluate the robustness of Galtier et al.’s results (1999).

**D3: Synthetic dataset with spatially homogeneous composition**

Replicates of D3 are generated using a TN93 substitution model (see model section above) along the branches of a 10-species tree. Arbitrary but reasonable values are chosen for the branch lengths (see figure 1). The substitution model is time-homogeneous and the base frequency distribution is expected to remain constant both across sites and lineages. Rate-heterogeneity is simulated using a gamma distribution with 20 equiprobable discrete gamma categories. The gamma shape parameter \( \alpha \) is set to 0.5, a reasonable value for the genes we are studying. This corresponds to an L-shaped distribution of the average substitution rate across sites with most of the sites evolving slowly. Exchangeabilities are set at reasonable values: \( \rho_{CT} = 2.5, \rho_{transversion} = 0.4 \) and the equilibrium frequencies are \( \{ \pi_A, \pi_C, \pi_G, \pi_T \} = \{0.40, 0.25, 0.15, 0.20 \} \).

**D4: Synthetic dataset with spatial compositional heterogeneity**

D4 is the equivalent of D3 with spatial compositional heterogeneity. The base frequency distribution is no longer assumed to be shared across sites and we use a different set of frequency parameters for each gamma category to simulate a correlation between composition and site-specific average substitution rate (see table 1 and the “real process” curve in figure 3). Fast-evolving sites are G+C rich and A+T poor compared to slow evolving sites. Frequency parameters for each gamma category were chosen so that the frequency distribution, averaged over the whole sequence, is that given for D3. Note that the process is still time-homogeneous and
stationary. Consequently, we expect the composition to be time-homogeneous for each gamma category and the pattern of compositional variation across sites is the same for contemporary and ancestral sequences.

**Empirical estimate of compositional variation across sites**

A likelihood-based method is used to estimate the site-specific average rate of evolution and to investigate associations between evolutionary rates and site-specific base composition. We neglect here the possibility of significant rate change at a site over evolutionary time (the covarion-hypothesis introduced by Fitch and Markowitz (1970)). Assuming a tree topology ($\tau$) and a gamma distribution of rates, we compute ML estimates (MLEs) for branch lengths ($\hat{\nu}$) and substitution model parameters ($\hat{\theta}$). These values maximize the probability that the assumed evolutionary model generated the observed sequences $P(X|\tau, \theta, \nu)$, computed by the pruning algorithm (Felsenstein, 1981). In an empirical Bayesian approach (Carlin and Louis, 2000; Yang and Wang, 1995), we use these MLEs to compute the posterior probabilities of each site $s$ being in a specific rate category $r_i$,

$$P(r_i|X_s, \tau, \hat{\theta}, \hat{\nu}) = \frac{P(X_s|r_i, \tau, \hat{\theta}, \hat{\nu}) \times P(r_i)}{\sum_{i=1}^{k} P(X_s|r_i, \tau, \theta, \nu) \times P(r_i)} ,$$

where $k$ is the number of rate categories and $X_s$ the data observed at column $s$ of the alignment. As mentioned before, rate categories have equal prior probability and $P(r_i) = 1/k$.

We are interested in the composition of each rate category and we estimate the distribution of frequencies conditional on rate using the posterior probabilities from equation 3 as weights. Then the expectation value of the frequency vector $\Pi$ conditional on rate $r_i$ is,

$$E(\Pi|r_i) = \frac{\sum_s E(\Pi|X_s)P(r_i|X_s)}{\sum_s P(r_i|X_s)} ,$$

(4)
where $E(\Pi|X_s)$ is approximated by the observed frequencies at each site.

**Modeling compositional heterogeneity across sites**

In this paper, we mention some issues related to the use of time-heterogeneous method with spatially heterogeneous datasets. To avoid confusion, we wish to emphasize that the new ASRV method proposed here is meant to model compositional variation across sites and not compositional variation in time. Actually, we expect this new model to behave badly in case of compositional biases among species.

The evolutionary processes that cause the observed compositional trends with respect to the site-specific evolutionary rate are poorly understood and, in any case, hard to model explicitly. One simple method is to use different equilibrium frequencies for each rate category and to consider each frequency vector as a free parameter to be estimated from the data. Nevertheless, this method has some disadvantages. It is parameter-rich and the number of gamma categories used has a direct impact on the model and the number of parameters (whereas discretization of the gamma distribution was primarily supposed to be a mathematical convenience). This substitution model was implemented and tested with simulated datasets and we found that the method did not fit the real frequency curves in practice.

We do not expect the composition to vary strongly between neighboring rate categories. Consequently, it seems reasonable to solve the issues with the previous unconstrained method by assuming that equilibrium frequencies vary smoothly with respect to the site-specific evolutionary rate. We tried simple parametric models to express frequencies as a function of the substitution rate but, again, we often found the fit unsatisfactory or limited to a specific dataset (work not shown).

We believe parametric methods failed because they were too rigid and we consequently used Gaussian processes to incorporate the prior of smoothness in our model of compositional variation. Gaussian processes are becoming popular in the machine learning community (MacKay, 1998) and they have already been applied
in various fields, usually for classification (Gibbs and MacKay, 2000) and regression problems (Williams and Rasmussen, 1996) (see also Chu et al. (2005) for an application in computational biology). Easy to implement and to interpret, Gaussian processes are also appealing because their nonparameteric nature does not constrain the model to a specific functional form. In our specific case, Gaussian processes can simply be considered as useful smoothing devices that returns a prior probability of observing a set of frequencies without making excessively strong assumptions on the underlying variation process. Frequencies at each rate category were treated as free parameters but strong variations were penalized by the Gaussian process prior. The method was implemented in the Bayesian framework using MCMC techniques.

Gaussian processes are fully defined by their mean and covariance function. The covariance matrix $C$ specifies how similar the composition is between each pair of rate categories. In phylogenetic inference, substitution rates are always expressed relative to each other and consequently the problem was reparameterized with $x_i = \log(r_i)$, where $r_i$ is the average substitution rate of the rate category $i$. We chose a covariance form adapted to our problem:

$$C(x_i, x_j) = \theta_0^2 \exp\left\{ -\theta_1^2(x_i - x_j)^2 \right\} + \theta_2^2 + \theta_3^2 x_i x_j + \delta_{ij} \theta_4^2. \tag{5}$$

The hyperparameter $\theta_0$ defines the amount of variation expected for a typical function, $\theta_1$ scales the rate abscissa, $\theta_2$ and $\theta_3$ introduce a linear trend in the Gaussian process and define the mean expected values of the process as a straight line with respect to the evolutionary rates $x$ while $\theta_4$ is a jitter element on the diagonal of the covariance matrix which prevents it from being ill conditioned. A distinct covariance matrix (i.e., a specific vector $\Theta_i = \{\theta_0, \theta_1, \theta_2, \theta_3, \theta_4\}$) is used for each of the $n$ possible states of the model (e.g., four nucleotides for TN93 and seven pairs for OTRNA) defining a set of hyperparameters $\Theta = \{\Theta_1, \Theta_2, \ldots, \Theta_n\}$. We used a unique (arbitrary) prior on each $\Theta_i$: $\theta_0 \sim E(10)$, $\theta_1 \sim E(1)$, $\theta_2 \sim U(0, 10)$,
\( \theta_3 \sim Un(0,10), \theta_4 = .005. \)

For each model, the \( n \) frequency parameters used at a specific rate category were reparameterized with \( n \) “activation” values using a softmax function to free ourselves from the usual constraints imposed on frequency parameters: \( 0 \leq \pi_i \leq 1 \) and \( \sum_i \pi_i = 1 \),

\[
\pi_i(r) = \frac{\exp a_i(r)}{\sum_j \exp a_j(r)}.
\] (6)

The \( a_i(x) \) vectors are not uniquely defined. It would have been possible to keep the extra degree of freedom for each gamma category and to let the Gaussian process prior resolve the identifiability issue, but we preferred to add the simple constraint \( \forall r, \sum_j a_j(r) = 0 \), although it has some effects on the prior for the variations.

The probability density of a set of activations \( a_i = (a_i(r_1), \ldots, a_i(r_k)) \), where \( \{r_1, \ldots, r_k\} \) are the rates of the discrete gamma categories is:

\[
p(a_1, a_2, \ldots, a_n|\Theta_1, \Theta_2, \ldots, \Theta_n) = \frac{1}{Z} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} a_i^T C^{-1} a_i \right\} \prod_{j=1}^{k} \delta(\sum_{i=1}^{n} a_i(r_j)) \) (7)
\]

where \( C(i,j) = C(\log(r_i), \log(r_j)) \), \( Z \) is a normalizing factor and \( \delta(x) \) is the Dirac delta function.

When a standard substitution model is used for Bayesian inference in \textit{PHASE}, a Dirichlet proposal mechanism is used to propose new equilibrium frequencies from the current distribution vector during a MCMC run. Multiple frequency vectors are used in the substitution model presented here and we simply chose to update them independently using the same mechanism. Since strong correlations between the frequencies of neighboring categories were introduced, distant moves are now regularly refused and the parameters that control the spread of the Dirichlet proposals are initially chosen (and also tuned during the MCMC “burnin” period) to ensure a reasonable acceptance rate when we sample from the chain. Since large moves are not possible, mixing properties are not impressive and a huge number of cycles are necessary to obtain a reliable sample from the stationary distribution of each gamma category.
category. A possible solution would be to design a suitable move affecting all of the
frequencies at once in order to preserve the correlations but this problem was not
addressed here and we used very long runs instead. Two runs were performed for
each simulation and we checked that the final posterior probability distributions of
frequency parameters were similar for each gamma category.

Results

Correlation of substitution rate and composition

Using the empirical Bayesian method described above (equations 3 and 4), composi-
tional trends and their correlation with the site-specific rate of evolution were studied
in the primate dataset D1. As previously mentioned, a TN93 substitution model was
used with RNA loops and the OTRNA model was used with RNA stems. In both
models, a gamma distribution of rates was assumed and approximated with eight
discrete categories. The empirical method requires us to assume a specific topol-
yogy and we simply used the majority-rule consensus topology found by a Bayesian
analysis with this dataset (see Jow et al. (2002) for details on the algorithms and
MCMC proposals used in PHASE). In spite of some oddities (i.e., the unrealistic
gorilla-human clade as found in Hudelot et al. (2003)), we do not expect this choice
to affect significantly the results since substitution parameter estimates are often
observed to be quite robust to the assumed topology.

Graphs for the estimated composition of RNA loops and RNA helices in D1
are shown in figure 2. The estimated frequency distribution for each discrete rate
category is plotted against the average substitution rate of the category. In RNA
stems, frequencies of symmetrical pairs (e.g., $\pi_{G,U}$ and $\pi_{U,G}$) were summed for clarity.
In loops, we observe that $\pi_G$ is negatively correlated with the rate of evolution and
that $G$ is scarce at fast evolving sites. In RNA stems, we observe a striking increase
in $\pi_{A,U+U:A}$ and $\pi_{MM}$, correlated with the site-specific evolutionary rate, whereas
\( \pi_{G:C+G:C} \) is decreasing.

In order to ascertain whether this method is robust, it was tested with ten replicates of D3 (1,000 nucleotides) and ten replicates of D4 (20,000 nucleotides). Datasets were analyzed using a TN93 substitution model and assuming the true topology, known for these synthetic datasets. Rate heterogeneity was modelled with twenty discrete gamma categories in both cases. The test with the spatially heterogeneous dataset D4 was to evaluate the performance of the method. The test with the spatially homogeneous and small dataset D3 was done to ensure that we were not finding heterogeneity when there was none. Indeed, the Markov process is observed for a very short period of time and with limited sequence size and we wanted to test whether noise or other side-effects could be responsible for the observed trends with the real dataset D1.

Over the ten experiments with D3, frequencies estimated for each rate category were found to be quite close to the stationary distribution that was used to generate the sequences. Differences between the estimated composition for each category and the real uniform frequencies were in the tight range \([-2.48\%,+2.56\%]\) (95% confidence interval). Moreover, no significant trends related to the substitution rate were visible. Consequently, we rule out the possibility that a methodological bias is responsible for the trends observed with real datasets. Figure 3 shows results with dataset D4 (first replicate) which indicate that the method can identify compositional differences across sites but is probably underestimating the extent of variation when it exists.

**Effects on standard phylogenetic inference methods**

To assess the impact of spatial compositional heterogeneity in phylogenetic inference, one hundred replicates of D4 were generated (20,000 sites per alignment) and analyzed with a standard homogeneous substitution model. A TN93 substitution model with gamma distributed rate heterogeneity (twenty discrete rate categories) was used. This is the generative model for D3 but it cannot account for the variation
of composition across sites in D4. For each replicate, branch lengths and substitution model parameters (i.e., frequencies, exchangeabilities and gamma shape parameter) were reestimated by ML optimization, assuming the known tree topology.

The following values were recovered for the frequency parameters \{\pi_A = 38.7 \pm 0.5\%, \pi_C = 27.3 \pm 0.5\%, \pi_G = 16.3 \pm 0.4\%, \pi_T = 17.7 \pm 0.4\%\}. These MLEs were clearly biased towards the composition of fast evolving sites (given in table 1). The site-specific composition depends on the rate category with replicates of D4 but one would have expected the MLEs for the frequency parameters to be close to the empirical frequencies, averaged over the whole sequence, \{\pi_A = 40\%, \pi_C = 25\%, \pi_G = 15\%, \pi_T = 20\%\}. The bias also exists when ASRV is not modelled but it was less noticeable: \{\pi_A = 39.2 \pm 0.5\%, \pi_C = 25.4 \pm 0.4\%, \pi_G = 15.8 \pm 0.4\%, \pi_T = 19.6 \pm 0.4\%\}. Frequencies were not the only parameters affected. Yang et al. (1994) and Huelsenbeck and Nielsen (1999) noticed that branch lengths were underestimated when using a simpler evolutionary model that does not accommodate rate heterogeneity or transition/transversion rate variation across sites. Similarly, we noticed that all branch lengths were slightly underestimated (by 3% approximately) when spatial frequency variation was not accounted for. The estimation of exchangeability parameters was also affected, with \rho_{CT} = 2.02 \pm 0.15 and \rho_{transversion} = 0.37 \pm 0.02 instead of \rho_{CT} = 2.50 and \rho_{transversion} = 0.40. Experiments were repeated with replicates of D3 (same alignment size, 20,000 sites) to confirm that deviations from the expected result were only due to the compositional variation across sites in D4. When replicates of D3 were used, we naturally found that MLEs were close to their true values. Indeed, ML is consistent when the true model is used and the alignments were large enough to grant us such results. Since these different biases were not observed when replicates of dataset D3 were used, we can be quite confident that spatial compositional variation in the synthetic replicates D4 was responsible for these results.

Above, we presented results where the frequency distribution is a free parameter
of the model inferred with the other evolutionary parameters during ML optimization. Since stationarity is assumed, a sensible method, commonly applied in phylogenetic inference, could have been to estimate the frequency parameters directly from the empirical composition of the studied sequences (Whelan and Goldman, 1999). Experiments with D4 were repeated in such a setting and we found that biases were more pronounced when frequencies were fixed to their empirical values. Branch lengths were even more underestimated (by 4% approximately) as were the exchangeability parameters ($\rho_{CT} = 1.85 \pm 0.15$ and $\rho_{transversion} = 0.36 \pm 0.02$).

Effects on topology estimates were studied in the Bayesian framework. One hundred replicates of D4 were generated with 1,000 sites per alignment. We compared the Bayesian posterior probability (BPPs) of clades obtained when the true substitution model that generated D4 was assumed (i.e., base frequencies for each rate category were different and fixed to their true values), and when a homogeneous TN93 substitution model was assumed (i.e., with a unique equilibrium frequency vector estimated during the inference process). Twenty discrete gamma categories were used in both cases. On top of the biases highlighted above in the ML framework, we found that the BPPs used to measure the support for different clades were sensitive to the model used. Focusing on the clades with BPPs between 50% and 95% when the true model was used (the values that are usually reported on a consensus tree), corresponding BPPs obtained with the spatially homogeneous model were slightly (but significantly) different ($\pm 5\%$ on average, and $\pm 10\%$ to 20\% in some cases).

**Effects on time-heterogeneous methods**

Standard evolutionary models are stationary and time-homogeneous: frequency and exchangeability parameters are supposed to remain constant over time and over lineages and the distribution of character states is at equilibrium. Nevertheless, it is unrealistic to assume that the forces that drive evolution have been constant
over time and heterogeneous methods have been proposed (Yang and Roberts, 1995; Galtier and Gouy, 1998; Foster, 2004). Though details differ in the three works mentioned, the general idea is to use different homogeneous substitution models for each branch in the tree.

We used the program EVAL\_NH in the NHML package (Galtier et al., 1999), to analyze the effect of compositional variation across sites with time-heterogeneous methods. This program was chosen because the implemented substitution model has few parameters and is easily tractable. It has one equilibrium frequency, $\theta$, which is the equilibrium G+C content ($\theta/2 = \pi_G = \pi_C$ using our notation) and one transition/transversion exchangeability $\kappa$. The $\kappa$ parameter is shared all over the tree but $\theta$ is branch-specific and allowed to vary. With heterogeneous methods the likelihood is dependant on the root position (Yang and Roberts, 1995) and the ancestral G+C frequency ($\omega$) observed at the root node cannot be assumed to be equal to the stationary distribution of states anymore since the system is not at equilibrium. However, it is possible to infer it directly from the data during the inference process.

Replicates of D4 (100 alignments of 20000 nucleotides) were analyzed with EVAL\_NH. The true topology was assumed once again. Experiments were performed with and without a discrete gamma model, where eight categories were used in the first case. With the time-heterogeneous and rate-heterogeneous model implemented in NHML, we found that the inferred ancestral G+C frequency $\omega$ was 33.85\% (95\% CI:[33\%, 34.57\%]). Albeit the model that generated the replicates of D4 is spatially heterogeneous, it is still time-homogeneous and stationary so that the real ancestral G+C frequency in replicates of D4 is equal to the mean stationary frequency ($\pi_G + \pi_C = 40\%$) plus some stochastic noise. Nevertheless, the $\omega$ we found is clearly biased towards the G+C frequency of slow evolving sites ($\pi_G + \pi_C = 27.1\%$ for the slowest gamma category). The bias was less striking but still visible without rate-heterogeneity ($\omega=37.3\%, 95\%\ CI:[36.56\%, 38.06\%]$).
The model proposed by Galtier and Gouy (1998) is a time-heterogeneous version of the T92 model (Tamura, 1992). Since the T92 substitution model is less complex than the TN93 used to generate the data and since we used only eight discrete gamma categories to analyze datasets generated with twenty categories, one might object that the violation of these assumptions is responsible for the observed differences. Actually, when the experiments were repeated with replicates of D3, results confirmed that these assumptions had an impact on the results ($\omega = 38.8 \pm 0.7\%$ with rate heterogeneity among sites and $\omega = 39.7 \pm 0.8\%$ without). Nevertheless, deviations from the expected ancestral G+C frequency are not as great as deviations found with D4. Moreover, similar results were reproduced when the generative TN93 substitution model used to create the synthetic datasets D3 and D4 was replaced with a T92 model.

The G+C content of rRNA genes is known to be correlated to the optimal growth temperature in prokaryots (Galtier and Lobry, 1997). Galtier et al. (1999) analyzed the evolutionary history of the LSU and SSU genes (datasets similar to D2a and D2b) with the nonhomogeneous model described above in order to infer the ancestral G+C content, and the environmental temperature, of the LUCA. They found a low ancestral G+C content and naturally concluded that LUCA was a nonhyperthermophilic species. We tried to evaluate whether the effects of spatial-heterogeneity on time-heterogeneous methods could be responsible for this unexpected conclusion. Di Giulio (2000) already reported that Galtier et al. (1999)'s results could not be repeated using maximum parsimony and the bias we are reporting here is a likely explanation for this disagreement. Since the possible bias of the nonhomogeneous method is related to a directional trend in the composition with respect to the site-specific evolutionary rate, we studied the variations of the G+C content across sites in D2a and D2b as we did for D1. Loops and stems were not partitioned and we used a single TN93 substitution model (with 8 discrete rate categories) in the Bayesian empirical approach described previously. In D2a and D2b, the average G+C con-
Content is found to increase with the average evolutionary rate (see figure 4). A similar amount of variation is observed for both genes \(i.e., 10\%-15\%\). This suggests that the G+C content returned by the time-heterogeneous method is probably an underestimate in both cases. Note that the procedure we are using to estimate the correlation between site-specific rate and composition is assuming that the substitution process is homogeneous whereas the dataset D2 is clearly time-heterogeneous (and not only spatially heterogeneous). We do not expect the assumption of homogeneity to have a great impact on the site-specific rate estimation procedure and on the results of the Bayesian empirical method. Nevertheless, the curves for the two genes in figure 4 can only be considered as a mean across their respective dataset and the species-specific variation of G+C composition with respect to the evolutionary rate differs markedly from this average behavior.

**Modeling compositional heterogeneity across sites**

The new model that constrains compositional variations between neighboring rate categories with a Gaussian process was tested with dataset D1 and D4. The dataset D1 was analyzed using the same two substitution models as previously: TN93 for loops and OTRNA for stems. The eight gamma categories assumed in both models were assigned a different set of frequencies and Gaussian process priors were used in both models to avoid strong variations at neighboring rate categories. The synthetic dataset D4 was analyzed using a TN93 substitution model and eight gamma categories with variable frequencies.

Mean posterior estimates (MPEs) for the hyperparameters \(\Theta\) (equation 5) are given in table 2 for both runs. Figures 5 and 6 show the MPEs for the frequency parameters at each rate category. Inferred frequency parameters are compared: (i) with the frequency found with the Bayesian empirical method for the real primate dataset D1 (figure 5), (ii) with the stationary frequencies of the simulated process for heterogeneous dataset D4 (figure 6). The results in figure 6 show excellent agreement.
between inferred and true frequencies. The fit is less impressive in figure 5 (though reasonable) but one must keep in mind that the empirical estimates probably flatten the curves and underestimate the variation of frequencies across sites (as previously observed in figure 3).

This new model was also tested with a replicate of D3 (1,000 nucleotides) to perform a negative control. Though we recovered some trends, the real homogeneous frequency parameters are within the credibility intervals. Results are provided as supplementary material.

Discussion

Compositional trends in real sequences

The empirical Bayesian method that we proposed for the study of compositional variations among RNA sites suffers from some drawbacks. Mayrose et al. (2004) reported that Bayesian methods were performing quite well for the estimation of the site-specific substitution rate but we are still limited by the accuracy of this approach. One important issue is that the method is assuming compositional homogeneity across sites to classify sites according to their evolutionary rate and it is probably not appropriate when the objective is to study spatial frequency variations. Results with D4 (figure 3) suggest that the use of a pattern-homogeneous model and/or the use of rate category posterior probabilities as weights tend to flatten the resulting curves. Consequently, no claim is made concerning the accuracy of the curves plotted with this method and we believe that we might be underestimating the extent of spatial compositional variation in real data. However, though the method might not be reliable enough to quantify compositional variation numerically, put together with the negative results obtained with dataset D3, we think that the striking trends exhibited with real sequences are genuine and that spatial compositional variation is a common phenomenon that can be correlated with the
site-specific rate of evolution.

For the results with the primate dataset D1, we focus the discussion around the frequencies that vary across rate categories. Note that the high A content observed with D1 in unpaired RNA regions is explained by Gutell et al. (2000) who suggests that unpaired adenine nucleotides are crucial components for the formation of three-dimensional rRNA molecules. We previously argued that the biological processes responsible for compositional variations were hardly understood and difficult to model but we think the observed trends can mostly be explained by the combined effects of the mutation and selection processes. Presumably fast categories are under reduced selection and respond to the mutational pressure whereas the frequency distribution we observe at sites under purifying selection, reflects an average of the site-specific selection pressure over all slow evolving sites. This explanation fits nicely with the results. In loops, the decreasing trend in G content is certainly due to the strong mutational bias away from G. This mutational bias is often invoked to explain the low G content at the third codon position for mitochondrial protein-coding genes found on the H-strand (Gibson et al., 2005) and our results are consistent with the hypothesis of the deamination of C on the H strand which results in the decrease of G and increase of A in RNA products (Reyes et al., 1998). Indeed, the two rRNA genes, which account for two-thirds of D1, are on the H-strand too. Moreover, when D1 was split into separate datasets (tRNAs and rRNAs) and variation of composition was studied separately, the decreasing trend in G content is more striking for rRNAs than for tRNAs as was expected since tRNA genes can be found both on the L and the H strands (curves not shown). Nevertheless, since a consensus secondary structure was used, we cannot exclude the possibility of a “contamination” of the loop partition with slow evolving paired sites, more G+C rich than the standard composition in RNA loops. However, we would also expect strong compositional variations for the C and U bases in such a case and since G and A are the only unpaired nucleotides exhibiting trends, we find this explanation
Compositional trends in RNA helices are stronger and fit with what could reasonably be expected. The frequency of mismatches increases steadily with the evolutionary rate, reflecting a weaker selection pressure to maintain the RNA secondary structure at fast evolving sites. Since G:C pairs are thermodynamically stronger than A:U pairs and have a strong stabilizing effect on the structure of RNA molecules, it seems reasonable that slow evolving regions, subject to a stronger selection pressure, are G:C rich compared to fast-evolving regions. Nevertheless, the mutation bias away from G mentioned previously might also be responsible for the striking decrease in the G:C content in favor of the A:U content. More studies are probably needed to assess how ubiquitous the decrease in the G:C content is.

Our study here was limited to the correlations between site-specific evolutionary rate and composition in RNA genes. Though we did not test this assumption, similar trends certainly exist with protein-coding genes at the nucleotide/codon level and at the amino-acid level. Selective constraints imposed by the genetic code and mutational biases might give rise to unexpected patterns in other genes.

Impact on evolutionary estimates

When a single stationary distribution of frequencies is assumed and shared across sites evolving at different rates, an invariant site provides only one single independent sample of this distribution whereas, at the other extreme, a site with infinite substitution rate would provide \( N \) uncorrelated samples (\( N \) being the number of taxa in the alignment). Deviations of frequency estimates from the empirical composition of fast evolving sites consequently have a larger detrimental effect on the likelihood and, as confirmed by our results with simulated datasets, ML and Bayesian inference methods favor frequency parameters that are closer in composition to the fast evolving sites when they are estimated during the inference process. This was not tested here but we can reasonably expect the divergence between empirical and estimated
frequencies to grow as taxon sampling increases (because $N$ will increase). Though the effects of spatial compositional heterogeneity on the estimation of substitution parameters are worrying, the impact on branch lengths and topology estimates seems quite limited (but not negligible). Consequently, disregarding pattern-heterogeneity of the substitution process has probably little effect on the inferred phylogeny although we feel more studies are needed.

In this paper, we make the assumption that the spatial compositional variation observed in the studied sequences results from the variation of the substitution process across sites. This is a reasonable assumption for a set of closely related species but we must concede that compositional trends can also be generated (and consequently explained) by a process heterogeneous in time. Indeed, a time-heterogeneous method coupled with a standard model of ASRV can generate sequences with complex compositional variation patterns across sites since fast evolving sites react more quickly to the variations of the shared equilibrium distribution. The composition at fast and slow evolving site cannot remain homogeneous for long. Perhaps the main issue with Galtier and Gouy’s (1998) time-heterogeneous model is not that it does not take into account the possibility that the process is heterogeneous across sites but the fact that the composition of the ancestral sequence is assumed homogeneous.

The ancestral G+C frequency obtained when analyzing D4 with NHML is strongly biased towards the frequencies observed at slow evolving sites. Indeed, spatial variation of the G+C content, as seen in D4, can be accommodated by a time-heterogeneous process having a low ancestral G+C composition and a high G+C equilibrium frequency (as observed at fast evolving sites). The program NHML has been used in many studies so far. Effects of the bias are probably limited for phylogenetic reconstruction but we feel that studies that focused on the ancestral G+C estimate values should be complemented by a study of the correlation between the G+C content and the site-specific evolutionary rate to confirm the robustness of their results. Although we did not use exactly the same sequences and the same
alignment, the dataset used by Galtier et al. (1999) probably exhibited trends similar to what we found with D2 (see figure 4). If site-specific variation of the substitution process also has a role in the observed compositional heterogeneity across sites then it is quite likely that Galtier et al. (1999) underestimated the actual ancestral G+C content of the two rRNA genes. The spatial G+C trend estimated in D2 is comparable, in range and direction, to the G+C variation we observed when D4 was analyzed with the empirical method (see above and figure 3). The number of species, shape of the tree and branch lengths certainly have a big influence on the extent of the underestimation and, more importantly, the real substitution process acting with D2 is probably both heterogeneous in time and in space unlike the simulated process that generated D4. We do not think that the extent of underestimation with D2 and D4 is similar but we cannot help noticing that if both ancestral G+C contents of the two real rRNA genes have been underestimated by as much as 6% (as was found with D4), then it would almost be possible to conclude that LUCA lived in a thermophilic environment. We think that a model combining both variation in time and variation across sites is necessary for a more accurate estimate of the ancestral composition.

Model of spatial heterogeneity

In Thorne et al. (1996) and in subsequent works from the Goldman group (e.g., Goldman et al. (1998); Liò and Goldman (2002)), a trained HMM model is used to incorporate information on the variation of the substitution patterns across sites before the inference. On the contrary, in more recent works (Pagel and Meade, 2004; Lartillot and Philippe, 2004; Soyer et al., 2002), evolutionary models that could potentially detect and characterize new patterns of heterogeneity across sites with little a priori knowledge were used. Incorporating strong prior constraints to prevent variations at neighboring rate categories clearly prevents the putative patterns of evolution to emerge freely in our model. Even though different substitution models
are used for each rate category, we did not completely relax the “one model fits all” assumption used in standard substitution models.

The mutational process can be assumed constant across sites and we believe it is fine to suppose that frequencies at fast evolving categories are correlated. Nevertheless, one can reasonably argue that slow evolving sites should not be grouped according to their average substitution rate but rather according to the site-specific selection pressure which drives the process equilibrium frequencies. While we agree with that point, we wish to emphasize that the new model we propose is primarily designed for RNA helices. We believe that slow evolving pairs can be treated with the same process since structural stability is the common and predominant selection factor.

Results using a Gaussian process prior on the primate dataset D1 and the synthetic dataset D4 are promising. With D1, inferred frequencies do not seem to fit very well the empirical curves (figure 5) but, as previously mentioned, results from the empirical Bayesian methods are just a best guess derived from a homogeneous model and the actual amount of variation might be seriously underestimated. Results with D4 are quite impressive (figure 6) because the alignment is of sufficient size and because the frequencies of the D4 process were close to linear in the logged substitution rate scale which is an assumption of the prior. Sensitivity and effect of the prior should always be investigated in Bayesian inference. In this case, frequencies inferred at slow evolving sites are clearly more influenced by the trend imposed by the prior rather than the data. A low value for the hyperparameter $\theta_1$ is important to prevent the unrealistic variations observed with the unconstrained method at slow evolving sites. With some datasets we found that the method would not behave properly unless $\theta_1$ was constrained close to 0 with an exponential prior. One obvious reason might be that patterns uncorrelated to the evolutionary rate were present in these real datasets and negating our attempt to smooth variations. However, we believe that time-heterogeneity and “difficult” species are a more likely
explanation for these occasional failures. More work is clearly needed to understand and evaluate the impact of the Gaussian process prior used to smooth frequency variations.

Conclusion

We have found a heterogeneous base composition in RNA genes that is most likely due to variation of the substitution process across sites. The distribution of states observed at a site is correlated to the site-specific evolutionary rate and a new method is proposed to capture this aspect of sequence evolution. Variation of equilibrium frequencies across sites is not accounted for by most evolutionary models and we investigated the effects of spatial compositional variation with standard methods. Impact on phylogenetic reconstruction is probably limited but we demonstrate that frequency parameter estimates in statistical phylogenetic inference methods are biased towards the frequencies of rapidly evolving sites. With nonhomogeneous methods, ancestral frequency estimates are biased towards frequency distributions of conserved sites and we recommend caution in applying these methods.

Acknowledgments

We thank anonymous reviewers for helpful comments and valuable suggestions. VGS gratefully acknowledges a studentship from the School of Computer Science at the University of Manchester.
Literature Cited

Brooks, D. J., J. R. Fresco, and M. Singh

Bruno, W. J.

Carlin, B. P. and T. A. Louis

Chu, W., Z. Ghahramani, F. Falciani, and D. L. Wild

Di Giulio, M.

Di Giulio, M.

Dimmic, M. W., D. P. Mindell, and R. A. Goldstein

Felsenstein, J.

Felsenstein, J.  
2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies.  

Felsenstein, J.  

Fitch, W. M. and E. Markowitz  

Foster, P. G.  

Galtier, N. and M. Gouy  

Galtier, N. and J. R. Lobry  

Galtier, N., N. Tourasse, and M. Gouy  
1999. A nonhyperthermophilic common ancestor to extant life forms.  

Gibbs, M. N. and D. J. C. MacKay  
*IEEE-NN*, 11:1456.
Gibson, A., V. Gowri-Shankar, P. G. Higgs, and M. Rattray

Goldman, N., J. L. Thorne, and D. T. Jones


Halpern, A. L. and W. J. Bruno

Hudelot, C., V. Gowri-Shankar, H. Jow, M. Rattray, and P. Higgs

Huelsenbeck, J. P. and R. Nielsen

Jermiin, L. S., S. Y. W. Ho, F. Ababneh, J. Robinson, and A. W. D. Larkum

Jow, H., C. Hudelot, M. Rattray, and P. G. Higgs
Kimura, M. and T. Otha

Lartillot, N. and H. Philippe

Liò, P. and N. Goldman

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny

MacKay, D. J. C.

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko

Nielsen, R. and Z. Yang

Pagel, M. and A. Meade
Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa

Reeves, J. H.

Reyes, A., C. Gissi, G. Pesole, and C. Saccone

Savill, N. J., D. C. Hoyle, and P. G. Higgs

Schwartzman, D. W. and C. H. Lineweaver

Seo, T., H. Kishino, and J. L. Thorne

Soyer, O., M. W. Dimmic, R. R. Neubig, and R. A. Goldstein

Stephan, W.
Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis

Tamura, K.

Tamura, K. and M. Nei

Thorne, J. L., N. Goldman, and D. T. Jones

Tillier, E. R. M. and R. A. Collins

Whelan, S. and N. Goldman

Williams, C. K. I. and C. E. Rasmussen

Wuyts, J., P. De Rijk, Y. Van de Peer, T. Winkelmans, and D. W. R.
Wuyts, J., Y. Van de Peer, T. Winkelmans, and D. W. R.

Yang, Z.


Yang, Z.


Yang, Z.


Yang, Z., N. Goldman, and A. Friday


Yang, Z., R. Nielsen, G. N., and A. M. Krabbe Pedersen


Yang, Z. and D. Roberts


Yang, Z. and T. Wang

Table 1

Frequency parameters used to generate D4.

<table>
<thead>
<tr>
<th>rate category</th>
<th>substitution rate</th>
<th>$\pi_A$</th>
<th>$\pi_C$</th>
<th>$\pi_G$</th>
<th>$\pi_T$</th>
<th>$\pi_{C+G}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0013</td>
<td>0.4509</td>
<td>0.1875</td>
<td>0.0839</td>
<td>0.2778</td>
<td>0.2714</td>
</tr>
<tr>
<td>2</td>
<td>0.0092</td>
<td>0.4468</td>
<td>0.1957</td>
<td>0.0901</td>
<td>0.2674</td>
<td>0.2858</td>
</tr>
<tr>
<td>3</td>
<td>0.0251</td>
<td>0.4399</td>
<td>0.2035</td>
<td>0.0981</td>
<td>0.2585</td>
<td>0.3016</td>
</tr>
<tr>
<td>4</td>
<td>0.0493</td>
<td>0.4327</td>
<td>0.2096</td>
<td>0.1077</td>
<td>0.2500</td>
<td>0.3173</td>
</tr>
<tr>
<td>5</td>
<td>0.0821</td>
<td>0.4256</td>
<td>0.2150</td>
<td>0.1171</td>
<td>0.2423</td>
<td>0.3321</td>
</tr>
<tr>
<td>6</td>
<td>0.1242</td>
<td>0.4187</td>
<td>0.2198</td>
<td>0.1258</td>
<td>0.2357</td>
<td>0.3456</td>
</tr>
<tr>
<td>7</td>
<td>0.1763</td>
<td>0.4135</td>
<td>0.2241</td>
<td>0.1330</td>
<td>0.2294</td>
<td>0.3571</td>
</tr>
<tr>
<td>8</td>
<td>0.2394</td>
<td>0.4086</td>
<td>0.2297</td>
<td>0.1395</td>
<td>0.2222</td>
<td>0.3692</td>
</tr>
<tr>
<td>9</td>
<td>0.3150</td>
<td>0.4038</td>
<td>0.2349</td>
<td>0.1458</td>
<td>0.2155</td>
<td>0.3807</td>
</tr>
<tr>
<td>10</td>
<td>0.4047</td>
<td>0.4003</td>
<td>0.2405</td>
<td>0.1514</td>
<td>0.2077</td>
<td>0.3919</td>
</tr>
<tr>
<td>11</td>
<td>0.5111</td>
<td>0.3961</td>
<td>0.2465</td>
<td>0.1573</td>
<td>0.2000</td>
<td>0.4038</td>
</tr>
<tr>
<td>12</td>
<td>0.6375</td>
<td>0.3934</td>
<td>0.2524</td>
<td>0.1622</td>
<td>0.1920</td>
<td>0.4146</td>
</tr>
<tr>
<td>13</td>
<td>0.7883</td>
<td>0.3893</td>
<td>0.2592</td>
<td>0.1681</td>
<td>0.1835</td>
<td>0.4273</td>
</tr>
<tr>
<td>14</td>
<td>0.9704</td>
<td>0.3858</td>
<td>0.2666</td>
<td>0.1733</td>
<td>0.1742</td>
<td>0.4399</td>
</tr>
<tr>
<td>15</td>
<td>1.1940</td>
<td>0.3819</td>
<td>0.2751</td>
<td>0.1786</td>
<td>0.1645</td>
<td>0.4537</td>
</tr>
<tr>
<td>16</td>
<td>1.4757</td>
<td>0.3772</td>
<td>0.2840</td>
<td>0.1841</td>
<td>0.1547</td>
<td>0.4681</td>
</tr>
<tr>
<td>17</td>
<td>1.8455</td>
<td>0.3716</td>
<td>0.2947</td>
<td>0.1894</td>
<td>0.1443</td>
<td>0.4841</td>
</tr>
<tr>
<td>18</td>
<td>2.3652</td>
<td>0.3646</td>
<td>0.3057</td>
<td>0.1941</td>
<td>0.1355</td>
<td>0.4998</td>
</tr>
<tr>
<td>19</td>
<td>3.2037</td>
<td>0.3564</td>
<td>0.3191</td>
<td>0.1978</td>
<td>0.1266</td>
<td>0.5169</td>
</tr>
<tr>
<td>20</td>
<td>5.5820</td>
<td>0.3429</td>
<td>0.3364</td>
<td>0.2027</td>
<td>0.1180</td>
<td>0.5391</td>
</tr>
</tbody>
</table>
Table 2

Mean posterior estimate for the hyperparameters of the Gaussian process prior

<table>
<thead>
<tr>
<th></th>
<th>$\theta_0$</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 - TN93 - A</td>
<td>0.0439</td>
<td>0.4438</td>
<td>3.3272</td>
<td>1.2966</td>
</tr>
<tr>
<td>D1 - TN93 - C</td>
<td>0.0419</td>
<td>0.4512</td>
<td>1.8451</td>
<td>2.2112</td>
</tr>
<tr>
<td>D1 - TN93 - G</td>
<td>0.0491</td>
<td>0.4180</td>
<td>3.5620</td>
<td>2.3813</td>
</tr>
<tr>
<td>D1 - TN93 - T</td>
<td>0.0476</td>
<td>0.3929</td>
<td>2.3093</td>
<td>2.0194</td>
</tr>
<tr>
<td>D1 - 7D - A:U</td>
<td>0.0575</td>
<td>0.4599</td>
<td>3.7055</td>
<td>2.3437</td>
</tr>
<tr>
<td>D1 - 7D - G:U</td>
<td>0.0607</td>
<td>0.5238</td>
<td>4.0645</td>
<td>2.3712</td>
</tr>
<tr>
<td>D1 - 7D - G:C</td>
<td>0.0649</td>
<td>0.4538</td>
<td>3.2261</td>
<td>2.5534</td>
</tr>
<tr>
<td>D1 - 7D - U:A</td>
<td>0.0643</td>
<td>0.4626</td>
<td>3.6084</td>
<td>1.9333</td>
</tr>
<tr>
<td>D1 - 7D - U:G</td>
<td>0.0596</td>
<td>0.4912</td>
<td>4.1492</td>
<td>2.6997</td>
</tr>
<tr>
<td>D1 - 7D - C:G</td>
<td>0.0631</td>
<td>0.4561</td>
<td>3.1251</td>
<td>2.1755</td>
</tr>
<tr>
<td>D1 - 7D - MM</td>
<td>0.0606</td>
<td>0.4663</td>
<td>2.4701</td>
<td>3.2818</td>
</tr>
<tr>
<td>D4 - TN93 - A</td>
<td>0.0569</td>
<td>0.4121</td>
<td>3.0336</td>
<td>1.8897</td>
</tr>
<tr>
<td>D4 - TN93 - C</td>
<td>0.0693</td>
<td>0.4270</td>
<td>2.3330</td>
<td>2.1618</td>
</tr>
<tr>
<td>D4 - TN93 - G</td>
<td>0.0570</td>
<td>0.4245</td>
<td>2.7306</td>
<td>2.1654</td>
</tr>
<tr>
<td>D4 - TN93 - T</td>
<td>0.0741</td>
<td>0.3885</td>
<td>2.7612</td>
<td>2.2992</td>
</tr>
</tbody>
</table>
Figure Legends

Fig. 1 – The synthetic tree used to generate datasets D3 and D4.

Fig. 2 – Frequency variation in mitochondrial RNA (a)loops and (b)helices for the primate dataset D1. Site-specific Bayesian posterior probabilities of rate categories and observed empirical composition were used to produce these plots.

Fig. 3 – Frequency variation in synthetic dataset D4. Same as Fig. 2 but the process generating the sequences is known in this case so we can evaluate the performance of the method.

Fig. 4 – Estimation of the G+C frequency variation with respect to the average evolutionary rate in the datasets D2a (LSU rRNA) and D2b (SSU rRNA). Note that the SSU and LSU curves are not meant to be compared directly since the underlying topologies and branch lengths are not the same.

Fig. 5 – Frequency parameters inferred by the heterogeneous method for each gamma category with the real dataset D1, (a)loops and (b)stems. The 95% credibility intervals around the mean posterior estimates are shown. Results with the empirical Bayesian method are superimposed for comparison.

Fig. 6 – Frequency parameters inferred by the heterogeneous method for each gamma category with the synthetic dataset D4. The 95% credibility intervals around the mean posterior estimates are shown. Stationary equilibrium frequencies of the real process are superimposed for comparison. Twenty discrete gamma categories were used to generate the sequences but only eight were used during the inference and therefore the discrete gamma model of the real process spans a larger range of rates.
Figure 1: The synthetic tree used to generate datasets D3 and D4.

Figure 2: Frequency variation in mitochondrial RNA (a) loops and (b) helices for the primate dataset D1. Site-specific Bayesian posterior probabilities of rate categories and observed empirical composition were used to produce these plots.
Figure 3: Frequency variation in synthetic dataset D4. Same as Fig. 2 but the process generating the sequences is known in this case so we can evaluate the performance of the method.

Figure 4: Estimation of the G+C frequency variation with respect to the average evolutionary rate in the datasets D2a (LSU rRNA) and D2b (SSU rRNA). Note that the SSU and LSU curves are not meant to be compared directly since the underlying topologies and branch lengths are not the same.
Figure 5: Frequency parameters inferred by the heterogeneous method for each gamma category with the real dataset D1, (a) loops and (b) stems. The 95% credibility intervals around the mean posterior estimates are shown. Results with the empirical Bayesian method are superimposed for comparison.

Figure 6: Frequency parameters inferred by the heterogeneous method for each gamma category with the synthetic dataset D4. The 95% credibility intervals around the mean posterior estimates are shown. Stationary equilibrium frequencies of the real process are superimposed for comparison. Twenty discrete gamma categories were used to generate the sequences but only eight were used during the inference and therefore the discrete gamma model of the real process spans a larger range of rates.