Unbiased estimate of synonymous and non-synonymous substitution rates with non-stationary base composition: supplementary material

Laurent Guéguen and Laurent Duret

1 Mathematics

1.1 Ability

Here we describe how stochastic mapping is used to compute the ability of a model for a set of sequences along a phylogenetic tree.

We denote by \( D \) a set of sequences in an alphabet \( A \), and by \( T \) a phylogenetic tree. On each site, each sequence of \( D \) is the result of a substitution process from a root sequence along the branches of \( T \). On a given branch \( b \) of length \( t \), this substitution process can be represented by a continuous time Markov process \( (X(\tau))_\tau \).

We define \( E = \{(a', a) \in A \times A; a \neq a'\} \) the set of all substitutions, and focus on a subset \( L \) of events \((L \subset E)\). These events are named \( L \)-events. In our case, the \( L \)-events are the synonymous substitutions or the non-synonymous substitutions and \( A \) is the set of codons.

\( N_L \) denotes the number of \( L \)-events that occur along process \( X \). Since \( X \) is unknown, \( N_L \) is unknown. Substitution mapping approach is used to compute the expectation of \( N_L \) over the distribution of \( X \), given branch \( b \), model \( \mathcal{M} \), and data \( D \), i.e. \( E(N_L | b, \mathcal{M}, D) \).

Now we define and compute the ability of a model \( \mathcal{M}' \) (with generator \( Q' \)) along this process \( X \).

At time \( \tau \), during a short time \( d\tau \), we define the instantaneous ability of \( \mathcal{M}' \) to perform \( L \)-event, \( A^{Q'}_L(\tau) \), as the expectation - on \( X(\tau) \) - of the number of \( L \)-events that would have been performed by a process following model \( \mathcal{M}' \) during \( d\tau \):

\[
A^{Q'}_L(\tau) = \sum_{(a, a') \in L} Q'_{a, a'}P(X(\tau) = a)d\tau
= \sum_{a \in A} Q'_{a, L}P(X(\tau) = a)d\tau
\]

where \( Q'_{a, L} = \sum_{a' \in A; (a, a') \in L} Q'_{a, a'} \).

The ability \( A^{Q'}_L \) is the mean value of this sum along the process \( X \):
\[ A_{L}^{Q'} = \frac{1}{t} \int_{\tau=0}^{t} A_{L}^{Q'}(\tau) d\tau \]

\[ = \frac{1}{t} \int_{\tau=0}^{t} \sum_{a \in A} Q'_{a,\perp} P(X(\tau) = a) d\tau = \frac{1}{t} \sum_{a \in A} Q'_{a,\perp} \int_{\tau=0}^{t} P(X(\tau) = a) d\tau \]

\[ = \frac{1}{t} \sum_{a \in A} Q'_{a,\perp} T_a \]

where \( T_a = \int_{\tau=0}^{t} P(X(\tau) = a) d\tau \) is the time spent by \( X \) in state \( a \).

As for stochastic mapping, the expectation of \( A_{L}^{Q'} \) over all \( X \),

\[ E(A_{L}^{Q'} | b, M, D) \]

can be efficiently computed. In Minin and Suchard (2008), \( t.A_{L}^{Q'} \) is defined as the reward of vector \( Q'_{a,\perp} = (Q'_{a,\perp})_a \), which expectation over all scenarios given branch \( b \), model \( M \), and data \( D \), can be computed in the same way as \( E(N_{L} | b, M, D) \).

### 1.2 Estimates of \( dN \) and \( dS \)

Here, we show that the proposed estimates of \( dN \) and \( dS \), using the ability of a neutral model, are the most likely on a branch \( b \) of a tree \( T \), given a model \( M \) and data \( D \).

Given a model \( M' \) with generator \( Q' \), the log-likelihood of a process \( X \) along a branch \( b \) is:

\[ lL(X|M') = \sum_{a \in A} Q'_{aa} T_a + \sum_{(a,a') \in E} N_{aa'} \log(Q'_{aa'}) \]

where \( T_a = \int_{\tau=0}^{t} P(X(\tau) = a) d\tau \) is the time spent by \( X \) in state \( a \) and \( N_{aa'} \) is the number of substitutions from \( a \) to \( a' \) that occurred in \( X \) on branch \( b \). And we consider the expectation on the distribution of \( X \) given \( T \), \( M \) and \( D \) (we remove expectation condition \( |b, M, D \) for sake of clarity):

\[ E(lL|M') = \sum_{a \in A} E(T_a).Q'_{aa} + \sum_{(a,a') \in E} E(N_{aa'}) \log(Q'_{aa'}) \]

\[ = -\sum_{(a,a') \in E} E(T_a).Q'_{aa'} + \sum_{(a,a') \in E} E(N_{aa'}) \log(Q'_{aa'}) \]

Now, we look for model \( M' \) that maximizes this likelihood. Actually, we only focus on the factors that define non-neutrality, i.e. the factors that discriminate synonymous substitutions from non-synonymous substitutions. We take into consideration two sets of substitutions : \( S \) (resp. \( N \)) the set of synonymous (resp. non-synonymous) substitutions. \( S \cup N = E \).

And we consider that \( Q' \) can be written as :

\[ Q'_{aa'} = \begin{cases} 
\alpha Q'_{aa'}^0 & \text{if } (a,a') \in S \\
\beta Q'_{aa'}^0 & \text{if } (a,a') \in N 
\end{cases} \]

where \( Q'_{aa'}^0 \) does not depend on the synonymous property of the substitution from \( a \) to \( a' \) (it is the “neutral” part of \( Q' \), a part that we do not want to estimate).

Then:
\[ E(\{L|\mathcal{M}\}) = - \sum_{(a,a') \in \mathcal{S}} E(T_a) \alpha Q^0_{aa'} - \sum_{(a,a') \in \mathcal{N}} E(T_a) \beta Q^0_{aa'} \\
+ \sum_{(a,a') \in \mathcal{S}} E(N_{aa'}) \log(\alpha Q^0_{aa'}) + \sum_{(a,a') \in \mathcal{N}} E(N_{aa'}) \log(\beta Q^0_{aa'}) \\
= -\alpha tE(A_S^0) - \beta tE(A_N^0) \\
+ \log(\alpha)E(N_\mathcal{S}) + \log(\beta)E(N_\mathcal{N}) + \sum_{(a,a') \in \mathcal{S}} E(N_{aa'}) \log(\alpha Q^0_{aa'}) \\
+ \sum_{(a,a') \in \mathcal{N}} E(N_{aa'}) \log(\beta Q^0_{aa'}) \]

with \( A^0 := A^{Q^0} \).

Now, we look for which values of \( \alpha \) and \( \beta \) \( E(\{L|\mathcal{M}\}) \) is maximized:

\[ \frac{\partial E(\{L|\mathcal{M}\})}{\partial \alpha} = -tE(A_S^0) + \frac{E(N_\mathcal{S})}{\alpha} = 0 \iff \alpha = \frac{E(N_\mathcal{S})}{tE(A_S^0)} \]

\[ \frac{\partial E(\{L|\mathcal{M}\})}{\partial \beta} = -tE(A_N^0) + \frac{E(N_\mathcal{N})}{\beta} = 0 \iff \beta = \frac{E(N_\mathcal{N})}{tE(A_N^0)} \]

Finally, given the fixed neutral part \( Q^0 \) (and given \( T, \mathcal{M}, D \)), the most-likely model on branch \( b \) is:

\[ Q'_{aa'} = \begin{cases} 
\frac{E(N_\mathcal{S})}{E(A_S^0)} Q^0_{aa'} & \text{if } (a,a') \in \mathcal{S} \\
\frac{E(N_\mathcal{N})}{E(A_N^0)} Q^0_{aa'} & \text{if } (a,a') \in \mathcal{N}
\end{cases} \]

and the most likely estimator of \( \omega \) is \( \frac{E(N_\mathcal{S})}{E(A_S^0)} \frac{E(A_N^0)}{E(N_\mathcal{N})} \).

\( dN \) and \( dS \) are usually defined as the (non-)synonymous numbers of substitutions per (non-)synonymous nucleotide. In order to fit with this definition, since \( Q' \) is normalized to perform one substitution per codon and per unit of time on a sequence at equilibrium, the computed estimates have to be divided per 3 and multiplied by the length of the branch.

**References**

Minin, V. and Suchard, M. 2008. Fast, accurate and simulation-free stochastic mapping. *Phil. Trans. Roy. Soc. B*, 363: 3985–3995.
2 Figures

Figure S1: Estimates of $\omega = 0.1$ with (a) a stationary model and (b) a non-stationary model, on simulated data with changing G+C content. $\theta_{\text{root}}$: G+C frequency in the root sequence. $\theta_{eq}$: G+C equilibrium frequency of the simulation model.
Figure S2: Estimates of $dN$, $dS$ and $\frac{dN}{dS}$ with a stationary model (left) and non-stationary model (right), on simulated data with changing G+C content and $\omega = 0.9$. $\theta_{\text{root}}$: G+C frequency in the root sequence. $\theta_{\text{eq}}$: G+C equilibrium frequency of the simulation model.
Figure S3: Estimates of $dN$, $dS$ and $\frac{dN}{dS}$ with a stationary model (left) and non-stationary model (right), on simulated data with changing G+C content and $\omega = 1$. $\theta_{\text{root}}$: G+C frequency in the root sequence. $\theta_{\text{eq}}$: G+C equilibrium frequency of the simulation model.
Figure S4: **Estimates of** $dN$, $dS$ and $\frac{dN}{dS}$ **with a stationary model (left) and non-stationary model (right), on simulated data with changing G+C content and $\omega = 1.1$.** $\theta_{\text{root}}$: G+C frequency in the root sequence. $\theta_{\text{eq}}$: G+C equilibrium frequency of the simulation model.
Figure S5: **Estimates of** $dN$, $dS$ and $\frac{dN}{dS}$ **with a stationary model (left) and non-stationary model (right), on simulated data with changing G+C content and $\omega = 2$.** $\theta_{\text{root}}$: G+C frequency in the root sequence. $\theta_{\text{eq}}$: G+C equilibrium frequency of the simulation model.
Figure S6: Estimates of \( dN \), \( dS \) and \( \frac{dN}{dS} \) with a stationary model (left) and non-stationary model (right), on simulated data with changing G+C content in codon position 1, and \( \omega = 0.1 \). \( \theta_{\text{root}} \): G+C frequency in codon position 1 of the root sequence. \( \theta_{\text{eq}} \): G+C equilibrium frequency in codon position 1 of the simulation model.
Figure S7: Estimates of \( dN \), \( dS \) and \( \frac{dN}{dS} \) with a stationary model (left) and non-stationary model (right), on simulated data with changing G+C content in codon position 2, and \( \omega = 0.1 \). \( \theta_{\text{root}} \): G+C frequency in codon position 2 of the root sequence. \( \theta_{\text{eq}} \): G+C equilibrium frequency in codon position 2 of the simulation model.
Figure S8: Estimates of \( dN \), \( dS \) and \( \frac{dN}{dS} \) with a stationary model (left) and non-stationary model (right), on simulated data with changing G+C content in codon position 3, and \( \omega = 0.1 \). \( \theta_{\text{root}} \): G+C frequency in codon position 3 of the root sequence. \( \theta_{\text{eq}} \): G+C equilibrium frequency in codon position 3 of the simulation model.
Figure S9: **Substitution rates estimated with codeml**. Sequences were simulated with changing G+C content and $\omega = 0.1$. From top to bottom: $dN$, $dS$ and $\frac{dN}{dS}$. $\theta_{\text{root}}$: G+C frequency in the root sequence. Dashed curve: estimates on sequences at equilibrium. $\theta_{\text{eq}}$: G+C equilibrium frequency of the simulation model.
Figure S10: \textbf{log2 of the ratios of estimates of} dN, dS \textbf{and} dN/dS with a stationary model over the estimates with a non-stationary model, in function of human GC3 content.
Figure S11: **Estimated difference between equilibrium GC3 in primate clade and root GC3** compared to observed human GC3.