Accurate estimations of evolutionary times in the context of strong CpG hypermutability

Mikael Falconnet*,

Institute for Evolution and Biodiversity,

Mathematical Institute,

Institute for Mathematical Statistics,

Westfälische Wilhelms-Universität,

Einsteinstr. 62, 48149 Münster, Germany,

phone: +49-251-83-33703, fax: +49-251-83-32712,

mikael.falconnet@uni-muenster.de

Sarah Behrens,

Institute for Evolution and Biodiversity,

Westfälische Wilhelms-Universität,

Hüfferstrasse 1, 48149 Münster, Germany

phone: +49-251-83-21096, fax: +49-251-83-24668,

s.behrens@uni-muenster.de

Running head: Evolutionary times in the context of CpG hypermutability

Key words: Markov model, Phylogenetic distances, DNA sequence evolution, neighbor-dependent substitution process, CpG hypermutability

*corresponding author

Abstract

We consider the substitution model T92+CpG of DNA sequence evolution which takes into account the hypermutability of CpG dinucleotides, an effect that can be especially observed in vertebrate genomes. We provide an exact method to simulate the evolution of finite DNA sequences under this model and numerical procedures to infer evolutionary times in two cases: between an ancestral and a present sequence and between two homologous sequences. We show on simulated data that our new numerical method yields very accurate estimations of divergence times. In a context of strong CpG hypermutability, it clearly outperforms the classical estimation procedure that is solely based on the model T92 without CpG influence.

1 Introduction

The calculation of distances for DNA sequences remains an important part of molecular phylogenetics. For instance, the widely used algorithm `PhyML` developed by Guindon and Gascuel (2003) designed to estimate large phylogenies by maximum likelihood starts with a fast distance-based tree. More importantly, it is a useful tool to point the strengths and weaknesses of Markov process models of nucleotide substitutions on which most of the phylogeny reconstruction methods are based. The present paper should be also seen in the light of this.

Most stochastic substitution models assume that nucleotide sites evolve independently from each other and, in general, according to a given Markovian kernel. Since the simplest model developed by Jukes and Cantor (1969), other phylogeneticists like Kimura (1980), Felsenstein (1981), Hasegawa *et al.* (1985), Tamura (1992), Yang (1994) etc. have incorporated more and more biological characteristics into substitution models to increase their accuracy. However, all these models still assume that there are no neighbor-dependencies between the different nucleotide sites. But this is actually not the case in a biological context. For instance, it is well known that in vertebrate species, the frequency of the dinucleotide CpG is much lower than the product of C and G frequencies (Bird (1980)). The CpG deficiency is a consequence of the methylation of cytosine in CpG dinucleotides since methylated CpG dinucleotides may change into TpG with higher frequency and consequently into CpA on the complementary strand. For example, in a large scale survey of the human genome, Wang *et al.* (1998) found that single-nucleotide polymorphisms are observed ten times more often at CpGs than at other dinucleotides. Thus, incorporating the influence of the neighborhood into substitution models seems to be inevitable.

This is the reason why Duret and Galtier (2000) introduced a model that superimposes neighbor-dependent substitutions CpG→CpA and CpG→TpG onto Tamura's substitution rates, both at the additional rate $r \geqslant 0$. Hereafter, this model will be denoted T92+CpG. However, introducing these neighbor-dependent substitution processes greatly complicates the inference of phylogenies and the estimation of divergence times: the distribution of a nucleotide at site i at a given time does not only depend on its value at previous times but also on the values of the surrounding sites $i-1$ and $i+1$ whose distributions in turn depend on the values at sites $i-2$, $i-1$, i, $i+1$ and $i+2$ at previous times and so on. Hence, one is faced with infinite-dimensional linear systems which are difficult to solve. To circumvent this problem, Duret and Galtier (2000) approximated trinucleotide frequencies (xyz) by dinucleotide frequencies (xy) in the following way: $(xyz) \approx (xy)(yz)/(y)$. Instead of relying on dinucleotide frequencies, Arndt and Hwa (2005) similarly assumed that word frequencies can be deduced from trinucleotide frequencies. However, all these approaches remain approximative and are not mathematically founded.

Entering the field of interacting particle systems, see Liggett (1985) for a review of the topic, Bérard *et al.* (2008) introduced a wide extension of the T92+CpG model of neighbor-dependent substitution processes, called RN+YpR, and investigated its theoretical properties. Thanks to the remarkable structural properties of this class of models, Bérard and Guéguen (2010) analyzed the performance of an original approach developed to show that the phylogenetic reconstruction methods in the context of independence between sites can be efficiently adapted to this class of models - within a mathematically rigorous framework and without resorting to approximations. They also compare their approach with the alternative strategy of Duret and Arndt (2008).

In the context of the Jukes Cantor model with CpG influence, i.e. the simplest but

non-trivial model of the RN+YpR model class, Falconnet (2010a) provided theoretical estimation procedures for the time elapsed between an ancestral sequence at stationarity and a present one and for the time of divergence between two homologous sequences which are originating from a common ancestral sequence at stationarity. For the ancestral case, Falconnet (2010a) showed that the result can be extended to the general RN+YpR models but he did not provide an explicit extension for the homologous case. The performance of these procedures has also not yet been investigated.

The goal of the present paper is to derive theoretical estimation procedures for homologous sequences in the T92+CpG model. Furthermore, assuming model T92+CpG, we develop numerical solutions adapted from Falconnet (2010a) to estimate the time elapsed both between an ancestral and a present sequence and between two homologous sequences. Our numerical estimation procedures are evaluated on simulated data and are compared to the classical estimation method by Tamura (1992) that is based on the T92 model without CpG influence. It turns out that the classical estimators for evolution times become inaccurate very fast when the influence of the CpG hypermutability increases. In contrast, our new estimators based on model T92+CpG remain robust as soon as the hypermutability is significant. Thus, considering that CpG hypermutability is one of the major determinants driving the evolution of human DNA sequences, we can conclude that our new estimation procedure clearly outperforms the classical one.

The paper is organized as follows. In Section 2, we describe the T92+CpG model, we explain how to simulate the evolution of DNA sequences under this model, we provide a method to infer phylogenetic distances for this neighbor-dependent model and we describe our simulation experiment. In Section 3, we present our main results on the comparison of the two procedures to estimate the divergence time of two homologous

DNA sequences. Section 4 is devoted to the discussion and impact of our findings. The Supplementary Material contains further mathematical elaboration related to Section 2 and additional tables for the simulation experiments.

2 Methods

2.1 Description of the T92+CpG model

The T92 + CpG model is a continuous-time model of DNA evolution where the sequences evolve under the combined effect of two superimposed mechanisms.

The first mechanism is an independent evolution of the sites as in the classical T92 model developed by Tamura (1992). Hence, it is characterized by a 4×4 matrix of substitution rates, each rate being the mean number of substitutions per unit of time. In our case, the matrix is defined as

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & (1-\theta)v & \theta v & \theta w \\ (1-\theta)v & \cdot & \theta w & \theta v \\ (1-\theta)v & (1-\theta)w & \cdot & \theta v \\ (1-\theta)w & (1-\theta)v & \theta v & \cdot \end{pmatrix} \end{array},$$

where v and w denote respectively the rates of transversional (T, C $\leftrightarrow$ A, G) and transitional (T $\leftrightarrow$ C or A $\leftrightarrow$ G) changes and θ denotes the G+C content.

A second mechanism is superimposed which describes the substitutions due to the influence of the neighborhood: the most noticeable case is based on experimentally observed CpG-methylation-deamination processes. Hence, we assume that the substitution rates of cytosine by thymine and of guanine by adenine in CpG dinucleotides are both increased by an additional nonnegative rate r.

This means that every C site whose right neighbor is not occupied by a G (every G site whose left neighbor is not occupied by a C resp.) changes at global rate $v+(1-\theta)w$,

i.e. after an exponentially distributed random time with mean $1/[v + (1-\theta)w]$, and when it does, it becomes an A (a T resp.) with probability $(1-\theta)v/[v + (1-\theta)w]$, a T (an A resp.) with probability $(1-\theta)w/[v + (1-\theta)w]$ and a G (a C resp.) with probability $\theta v/[v + (1-\theta)w]$. On the contrary, every C site whose right neighbor is occupied by a G (every G site whose left neighbor is occupied by a C resp.) changes at global rate $v + (1-\theta)w + r$, i.e. after an exponentially distributed random time with mean $1/[v+(1-\theta)w+r]$, and when it does, it becomes an A (a T resp.) with probability $(1-\theta)v/[v+(1-\theta)w+r]$, a T (an A resp.) with probability $[(1-\theta)w+r]/[v+(1-\theta)w+r]$ and a G (a C resp.) with probability $\theta v/[v + (1-\theta)w + r]$.

The case $r = 0$ corresponds to the classical T92 model. We recall that the T92+CpG model belongs to a larger class called RN+YpR introduced by Bérard *et al.* (2008) and we refer to this article for the main properties of the model.

2.2 Simulating the evolution of DNA sequences

Every model of the RN+YpR class can be constructed relying on the general principles based on infinitesimal generators described by Liggett (1985). However, Bérard *et al.* (2008) presented another construction based on Poisson processes called graphical representation, also developed by Liggett (1985), which is much more useful to simulate the evolution of a DNA sequence. We adapt this construction to our context.

In Falconnet (2010a), DNA sequences are presented as doubly infinite sequences of letters, i.e. elements of the set $\mathscr{A}^{\mathbb{Z}}$, where $\mathscr{A} = \{A, T, C, G\}$. Since real or simulated DNA sequences are obviously finite we need to explain how to simulate the evolution of a finite sequence under a model developed for doubly infinite sequences. Thanks to mathematical properties presented in Bérard *et al.* (2008), we know that the behavior of the sites in $\{1, \ldots, N\}$ is the same regardless of whether one considers that these sites

are embedded in a sequence indexed by $\mathbb{Z}$ or in a sequence indexed by $\{0, 1, \ldots, N, N+1\}$.

In our case, for a given sequence $X_{1:N} := (X_1, \ldots, X_N)$, where X_i is the value of the nucleotide at site i, we create a new sequence $X_{0:N+1}$ by adding two artificial nucleotides at sites 0 and $N+1$. We know from Bérard *et al.* (2008) that the choice of the nucleotides we add in 0 and $N+1$ will modify the evolution at the sites 0 and $N+1$ but not at the sites $\{1, \ldots, N\}$. With these two new nucleotides and the transformation of the interval as a circle, every site of the sequence possesses two neighbors.

It remains to describe how to simulate the evolution until a fixed time t of a circular sequence indexed by $\{0, 1, \ldots, N+1\}$. For every site i in $\{0, 1, \ldots, N+1\}$, we generate ten independent homogeneous Poisson processes until time t with the labels, rates and actions given in Table 1.

Table 1

Thus, we have $10 \times (N+2)$ independent homogeneous Poisson processes which run until time t. Let $\mathscr{T} = (t_k)_{k=0}^{K}$ denote the increasing list of random times generated by these random processes. We know that this list is almost surely finite, i.e. $K < \infty$, and that all the elements are almost surely pairwise distinct. Every time t_k in this list is associated with a site i_k and a label $\mathscr{L}_{i_k}^{x_k}$ with $\mathscr{L} \in \{\mathscr{U}, \mathscr{W}, \mathscr{R}\}$. Hence, to simulate the evolution of the sequence, it suffices to read the list from t_0 to t_K and execute the action associated with site i_k and label $\mathscr{L}_{i_k}^{x_k}$.

The theory of Poisson Processes and the paper from Bérard *et al.* (2008) suffice to prove that the simulation we provide matches the rates of the T92+CpG model.

2.3 Inferring distances and asymptotic confidence intervals

In the estimation procedures below, both constructions are based on dinucleotides encoded in the alphabet $\mathscr{B}$ defined as

$$\mathscr{B} = \{R, T, C\} \times \{Y, A, G\},$$

where R denote the set of purines defined as $\{A, G\}$ and Y the set of pyrimidines defined as $\{T, C\}$.

The estimators are based on various quantities provided by the alignment of the two sequences and we introduce some notations which are necessary for the rest of this section.

Notation 1. *For every t, $X(t)$ describes the whole sequence at time t and, for every i in $\mathbb{Z}$, the ith coordinate $X_i(t)$ of $X(t)$ is the random value of the nucleotide at site i and time t.*

For every ℓ and every word w of length $\ell+1$ written in the alphabet $\mathscr{A}$, we say that site i is occupied by w at time t if and only if $X_{i:i+\ell} = w$.

2.3.1 Ancestral case

The estimation procedure of the time elapsed between a present sequence and an ancestral one is based on the evolution of the quantity $(C, C)(t)$, where $(C, C)(t)$ denotes the frequency of sites occupied by C at time 0 (in the ancestral sequence) and by C at time t (in the present sequence) in doubly infinite DNA sequences.

Now, we recall from Falconnet (2010a) the definition of the estimator $(T_C^N)_{\text{est}}$ of the elapsed time and the main results about this estimator.

Definition 2. *Let $(T_C^N)_{\text{est}}$ denote the estimator of the elapsed time in the T92+CpG*

model defined as the solution in t of the equation

$$(C,C)(t) = (C,C)^N_{\text{obs}}, \tag{1}$$

where $(C,C)^N_{\text{obs}}$ denotes the observed value of $(C,C)(t)$ on aligned sequences of length N, i.e.

$$(C,C)^N_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{X_i(0) = X_i(t) = C\}.$$

Let $(\kappa^N_C)_{\text{obs}}$ and $(\nu^N_C)_{\text{obs}}$ denote the quantities defined as

$$\begin{aligned}(\kappa^N_C)_{\text{obs}} =& -\theta v(C,R)^N_{\text{obs}} - \theta w(C,T)^N_{\text{obs}} + [v + (1-\theta)w](C,C)^N_{\text{obs}} + r(C*,CG)^N_{\text{obs}},\\ (\nu^N_C)_{\text{obs}} =& (C,C)^N_{\text{obs}} - 5{(C,C)^N_{\text{obs}}}^2 + 2(CC,CC)^N_{\text{obs}} + 2(C*C,C*C)^N_{\text{obs}}.\end{aligned}$$

Then we have that

$$(\kappa^N_C)_{\text{obs}} \xrightarrow[N\to\infty]{a.s.} -(C,C)'(t).$$

It originates from the delta method used in Falconnet (2010a). The quantity $(\nu^N_C)_{\text{obs}}$ is the variance of $(C,C)^N_{\text{obs}}$.

Theorem 3 (Falconnet (2010a))**.** *Assume that the function $t \mapsto (C,C)(t)$ is a diffeomorphism and that the ancestral sequence is at stationarity. Then, when $N \to \infty$,*

$$(T^N_C)_{\text{est}} \xrightarrow{a.s.} t \quad \textit{and} \quad (\kappa^N_C)_{\text{obs}}\sqrt{N/(\nu^N_C)_{\text{obs}}}\left[(T^N_C)_{\text{est}} - t\right] \xrightarrow{d.} \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ denotes the standard normal law. An asymptotic confidence interval at

level ε for t is given by

$$\left[(T_C^N)_{\text{est}} - \frac{z(\varepsilon)}{(\kappa_C^N)_{\text{obs}}}\sqrt{\frac{(\nu_C^N)_{\text{obs}}}{N}}\ ,\ (T_C^N)_{\text{est}} + \frac{z(\varepsilon)}{(\kappa_C^N)_{\text{obs}}}\sqrt{\frac{(\nu_C^N)_{\text{obs}}}{N}}\right],$$

where $z(\varepsilon)$ denotes the unique real number such that $\mathbb{P}(|Z| \geqslant z(\varepsilon)) = \varepsilon$ with $Z \sim \mathcal{N}(0,1)$.

From Theorem 3, we can compute an estimator of the time elapsed provided that we are able to solve equation (1). To bypass this difficulty, we use the classical Runge-Kutta method (RK4). The RK4 method is an iterative method for the approximation of solutions of ordinary differential equations, see Butcher (2008) for more details. In Supplementary Material A.2.1 we explain the algorithm we use.

2.3.2 Homologous case

The estimation procedure of the time from divergence for two homologous sequences originating from a common ancestral sequence at stationarity is based on the evolution of the quantity $[C,C](t)$, where $[C,C](t)$ denotes the frequency of sites occupied by a C at time t in both doubly infinite present sequences. Similarly to the ancestral case, the definition of the estimator $[T_C^N]_{\text{est}}$ of the divergence time is as follows.

Definition 4. *Let $[T_C^N]_{\text{est}}$ denote the estimator of the elapsed time in the T92+CpG model defined as the solution in t of the equation*

$$[C,C](t) = [C,C]_{\text{obs}}^N, \tag{2}$$

where $[C,C]_{\text{obs}}^N$ denotes the observed value of $[C,C](t)$ of the aligned present sequences

X^1 *and* X^2 *of length* N*, i.e.*

$$[C, C]^N_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i^1(t) = X_i^2(t) = C\}.$$

Let $[\kappa_C^N]_{\text{obs}}$ *and* $[\nu_C^N]_{\text{obs}}$ *denote the quantities defined as*

$$\begin{aligned}
[\kappa_C^N]_{\text{obs}} =& - \sum_{(uv,xy) \in \mathscr{B} \times \mathscr{B}} Q_{\text{hom}}\left([uv, xy], [u'v', x'y']\right) [uv, xy]_{\text{obs}}, \\
[\nu_C^N]_{\text{obs}} =& [C, C]^N_{\text{obs}} - 5{[C, C]^N_{\text{obs}}}^2 + 2[CC, CC]^N_{\text{obs}} + 2[C * C, C * C]^N_{\text{obs}}.
\end{aligned}$$

The matrix Q_{hom} is defined in Supplementary Material A.2.2. Analogously to the ancestral case, we have that

$$[\kappa_C^N]_{\text{obs}} \xrightarrow[N \to \infty]{a.s.} -[C, C]'(t).$$

Theorem 5 (Falconnet (2010b))**.** *Assume that the function* $t \mapsto [C, C](t)$ *is a diffeomorphism and that the ancestral sequence is at stationarity. Then, when* $N \to \infty$*,*

$$[T_C^N]_{\text{est}} \xrightarrow{a.s.} t \quad \textit{and} \quad [\kappa_C^N]_{\text{obs}} \sqrt{N / [\nu_C^N]_{\text{obs}}} \left[[T_C^N]_{\text{est}} - t \right] \xrightarrow{d.} \mathscr{N}(0, 1),$$

where $\mathscr{N}(0, 1)$ *denotes the standard normal law. An asymptotic confidence interval at level* ε *for* t *is given by*

$$\left[[T_C^N]_{\text{est}} - \frac{z(\varepsilon)}{[\kappa_C^N]_{\text{obs}}} \sqrt{\frac{[\nu_C^N]_{\text{obs}}}{N}} \ , \ [T_C^N]_{\text{est}} + \frac{z(\varepsilon)}{[\kappa_C^N]_{\text{obs}}} \sqrt{\frac{[\nu_C^N]_{\text{obs}}}{N}} \right],$$

where $z(\varepsilon)$ *denotes the unique real number such that* $\mathbb{P}(|Z| \geqslant z(\varepsilon)) = \varepsilon$ *with* $Z \sim \mathscr{N}(0, 1)$*.*

Using Theorem 5, we can obtain an estimator of the divergence time of two homologous sequences if we are able to solve equation (2). In order to do so, as in the ancestral case, we use the classical Runge-Kutta method (RK4); see Supplementary Material A.2.2.

2.4 Simulation experiment: comparing estimation procedures

For all our simulations, we have focused on the T92+CpG model with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and several different values for the parameter r. The set of parameters (v, w, θ) is always assumed to be the same since we are only interested in the influence of the CpG hypermutability effect (specified by the parameter r) on the quality of the estimation procedures. The choice of parameters will be also discussed in Section 4.

In each case, the functions $t \mapsto (C, C)(t)$ and $t \mapsto [C, C](t)$ are diffeomorphisms as explained in Supplementary Material A.1. Thus, the conditions for Theorems 3 and 5 are fulfilled and the estimators $(T_C^N)_{\text{est}}$ and $[T_C^N]_{\text{est}}$ can be obtained by applying the Runge-Kutta method (RK4); see Supplementary Material A.2.1 and A.2.2.

Employing the tool "random sequence" from the RSAT suite with independent and equiprobable nucleotides, we generated random DNA sequences of length $N = 1000$, 10,000, 100,000 and 1,000,000 bp (Thomas-Chollier *et al.* (2008)). In order to obtain a DNA sequence at stationarity according to the T92+CpG model with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$, we wrote a *Python* script simulating the evolution of a given random sequence until time $t = 2$ as described in Section 2.2. This should suffice to produce stationary DNA sequences. As to be expected at stationarity, the resulting DNA sequences at time $t = 2$ contain 32.4% 'A's, 32.4% 'T's, 17.6% 'C's, 17.6% 'T's and the 'CpG' content is 0.96%. In the following, we only used these

stationary sequences as ancestral DNA sequences for the subsequent simulations.

Starting from every of these stationary ancestral DNA sequences, we then simulated its evolution until various time points t for different values of r, $r = 0, 0.1, 1, 2, 3, 4, 5, 10, 15$ and 20, according to the T92+CpG model with parameters $v = 1$, $w = 3$ and $\theta = 0.4$ (see Section 2.2). Each simulation has been carried out 100 times.

Our aim is to compare the results of our estimation procedure as described in Sections 2.3.1 and 2.3.2 to a classical estimation procedure of divergence times that is based on the T92 model without CpG influence by Tamura (1992) on the simulated data. The classical estimation procedure according to Tamura (1992) is described in Supplementary Material A.3. Let $(T^N_{\text{Tamura}})_{\text{est}}$ and $[T^N_{\text{Tamura}}]_{\text{est}}$ as defined in the Supplementary Material (see equations (6) and (7)) denote the classical estimators according to Tamura (1992) for the divergence time between an ancestral and a present sequence and between two homologous sequences respectiveley.

In the first scenario, we compared the evolved sequence $X(t)$ at time t, t ranging from 0.1 to 1.0, to the ancestral DNA sequence $X(0)$ and estimated the time elapsed between the two sequences using two different estimators: (i) our estimator $(T^N_C)_{\text{est}}$ based on model T92+CpG as described in Section 2.3.1, (ii) the classical estimator $(T^N_{\text{Tamura}})_{\text{est}}$ for model T92 according to Tamura (1992). In the second scenario, we let the ancestral sequence $X(0)$ evolve twice until time t and, for every time point t, t ranging from 0.1 to 1.0, compared the two resulting homologous sequences $X^1(t)$ and $X^2(t)$ to each other. As in the first scenario, we then estimated the divergence time between these two homologous sequences by (i) $[T^N_C]_{\text{est}}$ and by (ii) $[T^N_{\text{Tamura}}]_{\text{est}}$. In both scenarios, we additionally computed the 95% asymptotic confidence intervals of all the estimators based on the estimation method that has been employed, i.e. of $(T^N_C)_{\text{est}}$ and $[T^N_C]_{\text{est}}$ according to Theorems 3 and 5 and of $(T^N_{\text{Tamura}})_{\text{est}}$ and $[T^N_{\text{Tamura}}]_{\text{est}}$ according

to Supplementary Material A.3, equations (8) and (9). Finally, we checked whether the real t lies within this interval. We then used the percentage of real t lying in the confidence interval to assess the quality of the two estimation procedures.

When estimating the 95% asymptotic confidence interval for homologous sequences according to our estimation procedure, in some cases, especially when the sequence length N is small, $[\kappa_C^N]_{\text{obs}}$ can be negative resulting in a negative asymptotic variance (see Theorem 5). In this case the 95%-confidence interval is not defined. Thus, for further analyses we have filtered out these cases (for example, for $r = 10$ we had to filter out 25.4% of the estimations for $N = 1,000$ bp, 15.6% for $N = 10,000$ bp, 7.1% for $N = 100,000$ bp and only 0.5% for $N = 1,000,000$ bp). On the other hand, when estimating the divergence time between homologous sequences according to the classical estimation procedure, in some cases, especially when t is large, the estimator $[T^N_{\text{Tamura}}]_{\text{est}}$ is not well defined either (in equation (5), $1 - \frac{\hat{p}}{\theta_1+\theta_2-2\theta_1\theta_2} - \hat{q}$ can be negative). We have also filtered out these cases (for example, for $r = 10$, 31.3% of the estimations had to be filtered out for $N = 1,000$ bp, 16.4% for $N = 10,000$ bp, 9.9% for $N = 100,000$ bp and 2.5% for $N = 1,000,000$ bp).

3 Results

In this section, we assess the performance of the two distinct methods we consider for estimating evolutionary times, i.e. of our estimation procedure (see Sections 2.3 and 2.4) and of the classical estimation method based on the T92 model without CpG influence by Tamura (1992) (see Section 2.4 and Supplementary Material A.3). We mainly focus on the homologous case and only briefly deal with the ancestral case. We assess the impact of the CpG hypermutability effects on the accuracy of the estimation procedures and the impact of the sequence length on the size of the confidence interval computed.

3.1 Results in the ancestral case

Starting from a random stationary DNA sequence of length $N = 1,000,000$ bp, we simulated its evolution until time t, t ranging from 0.1 to 1.0, as described in Section 2.4 with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$. For every of the 100 simulatons, we estimated the real divergence time t between the ancestral and simulated sequence by $(T^N_{\text{Tamura}})_{\text{est}}$ (classical estimation) and by $(T^N_C)_{\text{est}}$ (new estimation). Additionally, we computed the 95% asymptotic confidence interval for every estimation and then examined if the real t lies in the 95% asymptotic confidence interval or not. Recall that the starting sequence is close to equilibrium for the T92+CpG model with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$. Hence, the conditions of Theorem 3 are fulfilled, and we expect the new procedure to be accurate according to the theory. For instance, since we carried out the simulation experiment 100 times, we expect the real t to be in the computed 95% confidence interval about 95 times. The results of this simulation experiment are depicted in Figure 1. As to be expected for the new estimation, in

Figure 1

most of the cases the real t lies in the computed confidence interval (more precisely: 100% for $t = 0.1$, 100% for $t = 0.2$, 99% for $t = 0.3$, 99% for $t = 0.4$, 100% for $t = 0.5$, 99% for $t = 0.6$, 98% for $t = 0.7$, 96% for $t = 0.8$, 98% for $t = 0.9$ and 94% for $t = 1.0$). In contrast, the real t never (i.e. for all t ranging from 0.1 to 1.0) lies in the 95% confidence interval of the classical estimation. As can be observed in Figure 1, the classical estimator drastically underestimates the divergence times. We can thus conclude that the classical estimation procedure is not suited for the estimation of the divergence time between an ancestral and a present DNA sequence whose evolution was influenced by CpG substitutions. For fairness, we have to mention that the starting random DNA sequence that we used for the simulation experiment is at stationarity according to the T92+CpG model. If we had started the simulations from a DNA sequence at stationarity according to the T92 model, the classical estimation would have probably yielded better results. However, assuming model T92 is less realistic. In order to alleviate the dependence on the starting DNA sequence, we decided to focus on comparing estimation results based on two homologous sequences that have diverged from their common ancestral sequence rather than reading too much into the comparison results for the ancestral case.

3.2 Results in the homologous case

3.2.1 Dependence on the CpG hypermutability effects

We examine how the accuracy of the classical and new estimation procedures for the divergence time of two homologous sequences changes when the CpG hypermutability effect varies, i.e. when the parameter r varies. All the simulations in this subsection have been carried out 100 times on the same random sequence of length $N = 1,000,000$ bp with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and for different values of r.

At first, like in the ancestral case, we let the starting sequence evolve with CpG hypermutability parameter $r = 10$ ending up with two homologous sequences for each divergence time t, t ranging from 0.1 to 1.0. We then estimated the divergence time t between the two homologous sequences by $[T^N_{\text{Tamura}}]_{\text{est}}$ (classical estimation) and by $[T^N_C]_{\text{est}}$ (new estimation) and calculated the corresponding 95%-confidence intervals. Since the starting sequence is close to equilibrium for the T92+CpG model with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$, the conditions of Theorem 5 are fulfilled and we expect the real t to be in the computed 95%-confidence interval in around 95% of the cases. The results are shown in Figure 2 and Supplementary Table 4. Indeed, the new procedure provides very accurate estimations ranging from 100% of real $t = 0.1$ being in the confidence interval to 89% of real $t = 1.0$ being in the confidence interval. The general tendency is that the new estimation procedure performs better for smaller than for larger divergence times t. Like in the ancestral case, the classical method underestimates the real t and its accuracy is very poor: for example, for t ranging from 0.1 to 0.7, the real t is never in the confidence interval, for $t = 0.8$ and $t = 0.9$ only in 1% of the cases and for $t = 1.0$ in 4% of the cases. The effect of a slightly higher accuracy for increasing divergence times t might be simply explained by random effects and by the fact that the length of the confidence interval increases with increasing divergence time t; see also Supplementary Table 6.

Figure 2

We then tested the effect of a varying CpG hypermutability parameter r on the accuracy of the estimations and repeated the simulation experiment for different values of r, r ranging from 0 (no CpG hypermutability effect) to 20 (strong CpG hypermutability effect). The outcome is depicted in Figure 3 and in Supplementary Tables 4 and 5. When the CpG hypermutability effect is absent, i.e. when $r = 0$, or when there is only a very weak effect, i.e. when $r = 0.1$, the classical estimation performs better than

Figure 3

the new method. But as soon as the CpG hypermutability effect increases or more precisely, when $r \geqslant 1$, the new estimation procedure outperforms the classical one. As can be seen in Supplementary Table 5 which gives the mean values of the estimated divergence times of the two methods, the stronger the CpG hypermutability effect gets, the more the classical method underestimates the real t. In constrast, the mean estimated divergence times of our method are always very close to the real t, sometimes slightly underestimating, sometimes slightly overestimating t. Furthermore, as a general tendency, we observe that for small r, the new estimation procedure improves with increasing divergence time t. Similarly, for large r, the classical estimation gets slightly better when t increases. But these effects can be simply explained by larger confidence intervals for increasing t that result in a higher probability of hitting the interval when estimating t; see also Supplementary Table 6. Generally, the new method reaches its optimal accuracy at around $r = 10$ while the classical procedure performs best for $r = 0$. However, contrary to the classical method which can basically just deal with cases where $r \leqslant 1$, the new estimation procedure can cope with a very broad range of CpG hypermutability parameters r and hardly looses its accuracy when r exceeds its optimum at 10.

3.2.2 Dependence on the sequence length

All the previous simulations have been carried out on a DNA sequence of length $N = 1,000,000$ bp. In order to test the impact of the sequence length on the accuracy of the new and classical estimation procedures, we repeated the simulation experiment for random DNA sequences (that are close to equilibrium for the T92+CpG model with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$) having different lengths N, N ranging from 10^3 to 10^6 bp. Like in the preceding subsection, we are also interested in the effect of the CpG hypermutability parameter r. Our simulation experiments

reveal that independently from the sequence length N, when $r \geqslant 1$, the new estimation procedure always outperforms the classical method as shown in Figure 4. The classical estimator is only more accurate when both $r = 0$ and $10^4 \leqslant N \leqslant 10^6$. Furthermore, we can observe that for $r \geqslant 1$, the accuracy of the new method increases as N increases. As opposed to this, the accuracy of the classical procedure decreases with increasing N. Thus, the difference in accuracy becomes most striking for $N = 10^6$: for long sequences that have evolved in the context of CpG hypermutability, i.e. with $r \geqslant 1$, our new method by far provides better estimations than the classical one.

Figure 4

We also tested how the sequence length influences the size of the 95%-confidence intervals of the two methods. The confidence interval lengths are illustrated in Figure 5. As can be observed, for $N \leqslant 10^5$ the new method produces larger confidence intervals than the classical procedure. However, for $N = 10^6$, the lengths of the two estimation procedures are almost the same. The mean lengths of the intervals for $N = 10^6$ are also given in Supplementary Table 6. Hence, applying the new estimation procedure instead of the classical method to sequences that are sufficiently large, i.e. larger than 1 Mb, does not result into significantly larger confidence intervals.

Figure 5

In sum, no matter how long the evolved sequences are, the new method produces more accurate results than the classical procedure as soon as the CpG hypermutability effects gets significant, i.e. when $r \geqslant 1$. However, in order to guaranty small asymptotic confidence intervals, the new method should be only applied for long sequences, i.e. sequences of around 1,000,000 bp or more.

4 Discussion

For the T92+CpG model, we have provided an exact method to simulate the evolution of finite DNA sequences and a numerical procedure to infer evolutionary times between two homologous sequences. We have shown on simulated data that this procedure is far more accurate than the usual ones in the context of strong CpG hypermutability. However, in order to guarantee small asymptotic confidence intervals it should be applied on long sequences, i.e. sequences of around 1 Mb or more.

Note that it is necessary to have a prior knowledge of the set of parameters (v, w, θ, r) to apply the procedure. When dealing with real instead of simulated DNA sequences, for this purpose, Bérard and Guéguen (2010) have developed a procedure to estimate parameters of the T92+CpG model which can be applied to real data and which has been shown to be accurate on simulated data. A combination of the two procedures should provide a powerful tool in the context of phylogenetic reconstruction, but has to be further investigated.

Another point which could be addressed in the future is the improvement of the quality of the estimators, particularly regarding the length of the confidence interval computed. In this paper, we provided an estimator $[T_C^N]_{\text{obs}}$ based on the alignment of cytosines, but we are able to provide an estimator $[T_x^N]_{\text{obs}}$ based on the alignment of nucleotides x for every $x \in \{A, T, C, G\}$. Hence, a convex combination of these four estimators is also a consistent estimator for evolutionary times. The idea would be to find the combination which provides the smallest variance and as a consequence, the smallest asymptotic confidence interval. The problem is that up to our knowledge, the variance cannot be computed easily and the problem has to be investigated carefully.

Finally, the main message of this work is that in the context of strong CpG hyper-

mutability, one should be careful with the use of independent models and be aware that the choice of such simplistic models can have a wide-ranging impact on the quality of evolutionary estimators. Our new estimation procedure incorporates CpG neighbor dependencies and thus, can be seen as a step towards more accurate evolutionary times estimators, especially in the context of vertebrate genomes exhibting a strong CpG hypermutability.

Acknowledgements

We would like to thank Grégory Vial, Leonor Palmeira, Laurent Guéguen and Kai Müller for helpful discussions.

Author Disclosure Statement

No competing financial interests exist.

References

Arndt, P. F. and Hwa, T., 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* 21, 2322–2328.

Bérard, J., Gouéré, J.-B., and Piau, D., 2008. Solvable models of neighbor-dependent nucleotide substitution processes. *Mathematical Biosciences* 211, 56–88.

Bird, A. P., 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research* 8, 1499–1504.

Butcher, J. C., 2008. *Numerical methods for ordinary differential equations.* John Wiley & Sons Ltd., Chichester, second edition.

Bérard, J. and Guéguen, L., 2010. Accurate phylogenetic estimation of substitution rates with context-dependent models. In preparation.

Duret, L. and Arndt, P. F., 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics* 4(5).

Duret, L. and Galtier, N., 2000. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Molecular Biology and Evolution* 17, 1620–1625.

Falconnet, M., 2010a. Phylogenetic distances for neighbour dependent substitution processes. *Mathematical Biosciences* 224(2), 101–108.

Falconnet, M., 2010b. *Sur deux problèmes mathématiques de reconstruction phylogénétique.* Ph.D. thesis, Université de Grenoble.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences : A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.

Guindon, S. and Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52, 696–704.

Hasegawa, M., Kishino, H., and Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174.

Jukes, T. H. and Cantor, C. R., 1969. *Mammalian protein metabolism*, chapter Evolution of Protein Molecules, 21–132. Academic Press, New York.

Kimura, M., 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. *J. Mol. Evol.* 10, 111–120.

Liggett, T. M., 1985. *Interacting Particle System*, *Grundlehren der Mathematischen Wissenschaften*, volume 276. Springer-Verlag, New York.

Tamura, K., 1992. Estimation of the number of nucleotide substitutions when there are strong transition/transversion and g+c content biases. *Molecular Biology and Evolution* 9, 678–687.

Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohée, S., and van Helden, J., 2008. RSAT: regulatory sequence analysis tools. *Nucleic Acids Research* 36, W119–W127.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D.,

Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander, E. S., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.

Yang, Z., 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39, 105–111.

A Supplementary Material

A.1 How to prove that $t \mapsto (C,C)(t)$ and $t \mapsto [C,C](t)$ are diffeomorphisms

Even if it is possible to compute an explicit formula for $(C,C)(t)$, it is not possible, up to our knowledge, to prove that $t \mapsto (C,C)(t)$ is a diffeomorphism for every set of parameters (v,w,θ,r). However, for every given set of parameters considered in this paper, we have an approximate formula for $(C,C)(t)$ that can be obtained by computing the exponential of the matrix Q_{anc} defined in Supplementary Material A.2.1.

For instance, when $v=1$, $w=3$, $\theta=0.4$ and $r=10$, computations with `Maple` reveal that $(C,C)(t)$ is clearly a diffeomorphism as can be seen in Figure 6. Figure 6

Analogously, the computation of the exponential of the matrix Q_{hom} defined in Supplementary Material A.2.2 gives an approximate formula for $[C,C](t)$.

A.2 The RK4 method

A.2.1 RK4 in the ancestral case

We explain in this section how to numerically solve equation (1) in t given by

$$(C,C)(t) = (C,C)^N_{\text{obs}}.$$

From Falconnet (2010a) we know that in the T92+CpG model, the evolution of dinucleotides encoded in the alphabet $\mathscr{B}$ is autonomous and Markovian and ruled by the 9×9 infinitesimal generator Q_{anc} given in Table 2. Table 2

Hence, if $U(t)$ denotes the time-dependent vector of length 9 defined as

$$U(t) = ((C*, uv)(t))_{uv \in \mathscr{B}},$$

where $(C*, uv)(t)$ denotes the frequency of sites occupied by $C*$ at time 0 and by uv at time t, $U(t)$ satisfies the initial value problem

$$U'(t) = f(t, U(t)), \quad U(0) = U_0, \tag{3}$$

where f is the function defined as

$$f(t, U) = Q_{\text{anc}}^T U \quad \text{with} \quad Q_{\text{anc}}^T \text{ being the tranposed matrix of } Q_{\text{anc}}$$

and U_0 the vector defined as

$$U_0 = (0, 0, 0, 0, 0, 0, (CY)_*, (CA)_*, (CG)_*)^T$$

with $(xy)_*$ being the stationary frequency of dinucleotide xy under the T92+CpG model.

With these notations, we have for every time t

$$(C, C)(t) = U(CY)(t) + U(CA)(t) + U(CG)(t). \tag{4}$$

Hence, if we are able to compute an approximate solution of the initial value problem (3), we get an approximation of the function $t \mapsto (C, C)(t)$ and consequently solve equation (1).

We know from Supplementary Material A.1 that the function $t \mapsto (C, C)(t)$ is a

decreasing diffeomorphism from $(C)_*$ to $(C)_*^2$ for the set of parameters we consider. It could happen that the observed value $(C,C)_{\text{obs}}^N$ does not belong to the interval $[(C)_*^2, (C)_*]$. In this case, it is not possible to compute a solution for t. However, we know that $(C,C)_{\text{obs}}^N$ converges almost surely to $(C,C)(t)$ when $N \to \infty$. Thus, if the length of the sequence is large enough, the probability of the event of $(C,C)_{\text{obs}}^N$ not belonging to the interval $[(C)_*^2, (C)_*]$ is vanishingly small. Note that to compute $(C)_*$, it suffices to compute $(CY)_*$, $(CA)_*$ and $(CG)_*$ which can be computed via the kernel of Q_{anc}^T.

We now assume that $(C,C)_{\text{obs}}^N \in [(C)_*^2, (C)_*]$. Thanks to the RK4 method as decribed, for example, by Butcher (2008), we are able to provide an algorithm to solve equation $(C,C)(t) = (C,C)_{\text{obs}}^N$ given a set of parameters (v,w,θ,r), an input value for $(C,C)_{\text{obs}}^N$ and a step h in Table 3. Table 3

Let n denote the number of times we iterate step 2 and $t_n = nh$. Let $(C,C)_n = U_n(CY) + U_n(CA) + U_n(CG)$ denote the approximation of $(C,C)(t_n)$. When the algorithm stops, one has

$$(C,C)(t_{n+1}) \approx (C,C)_{n+1} \leqslant (C,C)_{\text{obs}}^N < (C,C)_n \approx (C,C)(t_n).$$

Hence, the time t returned is an approximation of the solution of equation (1).

The RK4 method for problem (3) is specified by the following equations

$$U_{n+1} = U_n + \frac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4), \quad t_{n+1} = t_n + h,$$

with U_{n+1} being the RK4 approximation of $U(t_{n+1})$ and with

$$\begin{aligned} k_1 =& f(t_n, U_n), \\ k_2 =& f(t_n + \frac{1}{2}h, U_n + \frac{1}{2}hk_1), \\ k_3 =& f(t_n + \frac{1}{2}h, U_n + \frac{1}{2}hk_2), \\ k_4 =& f(t_n + h, U_n + hk_3). \end{aligned}$$

The error per step in the RK4 method is up to the order of h^5 but the total accumulated error is only on the order of h^4. This is sufficient for our problem.

A.2.2 RK4 in the homologous case

Solving equation (2) can be done analogously to the ancestral case. But instead of the 9×9 matrix Q_{anc}, one has to consider a 81×81 matrix Q_{hom}. We do not write the explicit infinitesimal generator Q_{hom} but we explain how to compute the entries of this matrix. We need to define the rate of change $[uv, xy] \to [u'v', x'y']$ for every uv, xy, $u'v'$, $x'y'$ in $\mathscr{B}$, where the state $[uv, xy]$ represents the situation where the site i is occupied by uv in the first sequence and by xy in the second sequence. This rate is given by the following formula:

$$Q_{\text{hom}}([uv, xy], [u'v', x'y']) = \begin{cases} Q_{\text{anc}}(uv, uv) + Q_{\text{anc}}(xy, xy) & \text{if } uv = u'v' \text{ and } xy = x'y', \\ Q_{\text{anc}}(uv, u'v') & \text{if } uv \neq u'v' \text{ and } xy = x'y', \\ Q_{\text{anc}}(xy, x'y') & \text{if } uv = u'v' \text{ and } xy \neq x'y', \\ 0 & \text{else.} \end{cases}$$

Equation (2) can then be solved using the RK4 method in analogy to the ancestral case replacing Q_{anc} by Q_{hom}, $U(t) = ((C*, uv)(t))_{uv \in \mathscr{B}}$ by $V(t) = ([uv, xy](t))_{uv, xy \in \mathscr{B}}$

and U_0 by $V_0 = ([uv, xy]_*)_{uv,xy\in\mathscr{B}}$.

A.3 Classical inference of distances and asymptotic confidence intervals in model T92

According to Tamura (1992), an estimator $\hat{d}$ of the time of divergence between two sequences in model T92 is given by

$$\hat{d} = -(\theta_1+\theta_2-2\theta_1\theta_2)\cdot\log\left(1-\frac{\hat{p}}{\theta_1+\theta_2-2\theta_1\theta_2}-\hat{q}\right)-\frac{1-\theta_1-\theta_2+2\theta_1\theta_2}{2}\cdot\log(1-2\hat{q}) \tag{5}$$

where $\hat{p}$ and $\hat{q}$ denote the estimators of the proportions of the nucleotide sites that show, respectively, transitional and transversional differences between the two sequences and where θ_1 and θ_2 denote the 'G+C' content in sequence 1 and sequence 2 respectively. When considering an ancestral $X(0)$ and a present sequence $X(t)$, p and q can be estimated by

$$\hat{p} = \frac{(C,T)^N_{\text{obs}} + (T,C)^N_{\text{obs}} + (A,G)^N_{\text{obs}} + (G,A)^N_{\text{obs}}}{N},$$

$$\hat{q} = \frac{\sum_{i=1}^N \mathbf{1}\{X_i(0)\neq X_i(t)\}}{N} - \hat{p}$$

where

$$(x,y)^N_{\text{obs}} := \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{X_i(0)=x,\ X_i(t)=y\},\quad x,y\in\mathscr{A}.$$

In the homologous case when comparing two present sequences $X^1(t)$ and $X^2(t)$, $\hat{p}$ and $\hat{q}$ are analogously given by

$$\hat{p} = \frac{[C,T]^N_{\text{obs}} + [T,C]^N_{\text{obs}} + [A,G]^N_{\text{obs}} + [G,A]^N_{\text{obs}}}{N},$$

$$\hat{q} = \frac{\sum_{i=1}^N \mathbf{1}\{X^1_i(t)\neq X^2_i(t)\}}{N} - \hat{p}$$

where

$$[x,y]^N_{\text{obs}} := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{X_i^1(t)=x,\ X_i^2(t)=y\}, \quad x,y \in \mathscr{A}.$$

If we want to compare the divergence time estimators from Tamura's classical estimation procedure to those from our estimation method, we have to account for the speed differences of the two different models T92 and T92+CpG. To do so, we have to divide $\hat{d}$ by the substitution rate λ_r per nucleotide per unit of time in the T92+CpG model which is given by:

$$\lambda_r = v + \theta w[(A)_* + (T)_*] + (1-\theta)w[(C)_* + (G)_*] + 2r(CG)_*,$$

where

$$\begin{aligned}(A)_* = (T)_* =& 1/4\frac{-r\theta^2 w - 6\theta v^2 - 2\theta w^2 - 8w\theta v - 4vr\theta + vr\theta^2 + 2rw + 6vr + 2w^2 + 8vw + 6v^2}{3v^2 + 4vw + vr\theta + 3vr + rw + w^2 + r\theta w},\\ (C)_* = (G)_* =& \theta/4\frac{6v^2 + 6vr + 8vw + 2rw + 2w^2 + r\theta w - vr\theta}{3v^2 + 4vw + vr\theta + 3vr + rw + w^2 + r\theta w}.\end{aligned}$$

Setting

$$(T^N_{\text{Tamura}})_{\text{est}} := \frac{\hat{d}}{\lambda_r} \tag{6}$$

for the ancestral case and

$$[T^N_{\text{Tamura}}]_{\text{est}} := \frac{\hat{d}}{2\lambda_r} \tag{7}$$

for the homologous case, we obtain modified classical estimators that can be compared to the equivalents $(T^N_C)_{\text{est}}$ and $[T^N_C]_{\text{est}}$ of our estimation procedure.

As shown by Tamura (1992), the variance $\text{var}(\hat{d})$ of the estimator $\hat{d}$ is given by

$$\text{var}(\hat{d}) = \frac{a^2\hat{p} + b^2\hat{q} - (a\hat{p} + b\hat{q})^2}{N}$$

where

$$a = \frac{1}{1 - \frac{\hat{p}}{\theta_1+\theta_2-2\theta_1\theta_2} - \hat{q}},$$

$$b = (\theta_1 + \theta_2 - 2\theta_1\theta_2) \cdot a + \frac{1 - \theta_1 - \theta_2 + 2\theta_1\theta_2}{1 - 2\hat{q}}.$$

Correcting again for the speed differences of model T92 and T92+CpG by dividing by λ_r, one obtains

$$\left[(T^N_{\text{Tamura}})_{\text{est}} - \frac{z(\varepsilon)}{\lambda_r}\sqrt{\frac{\text{var}(\hat{d})}{N}} \ , \ (T^N_{\text{Tamura}})_{\text{est}} + \frac{z(\varepsilon)}{\lambda_r}\sqrt{\frac{\text{var}(\hat{d})}{N}} \right] \tag{8}$$

as an asymptotic confidence interval at level ε for t in the ancestral case where $z(\varepsilon)$ denotes the unique real number such that $\mathbb{P}(|Z| \geqslant z(\varepsilon)) = \varepsilon$ with $Z \sim \mathscr{N}(0,1)$. Analogously,

$$\left[[T^N_{\text{Tamura}}]_{\text{est}} - \frac{z(\varepsilon)}{2\lambda_r}\sqrt{\frac{\text{var}(\hat{d})}{N}} \ , \ [T^N_{\text{Tamura}}]_{\text{est}} + \frac{z(\varepsilon)}{2\lambda_r}\sqrt{\frac{\text{var}(\hat{d})}{N}} \right] \tag{9}$$

is an asymptotic confidence interval at level ε for t in the homologous case.

A.4 Additional tables for the the simulation experiments

Table 4

Table 5

Table 6

List of Figures

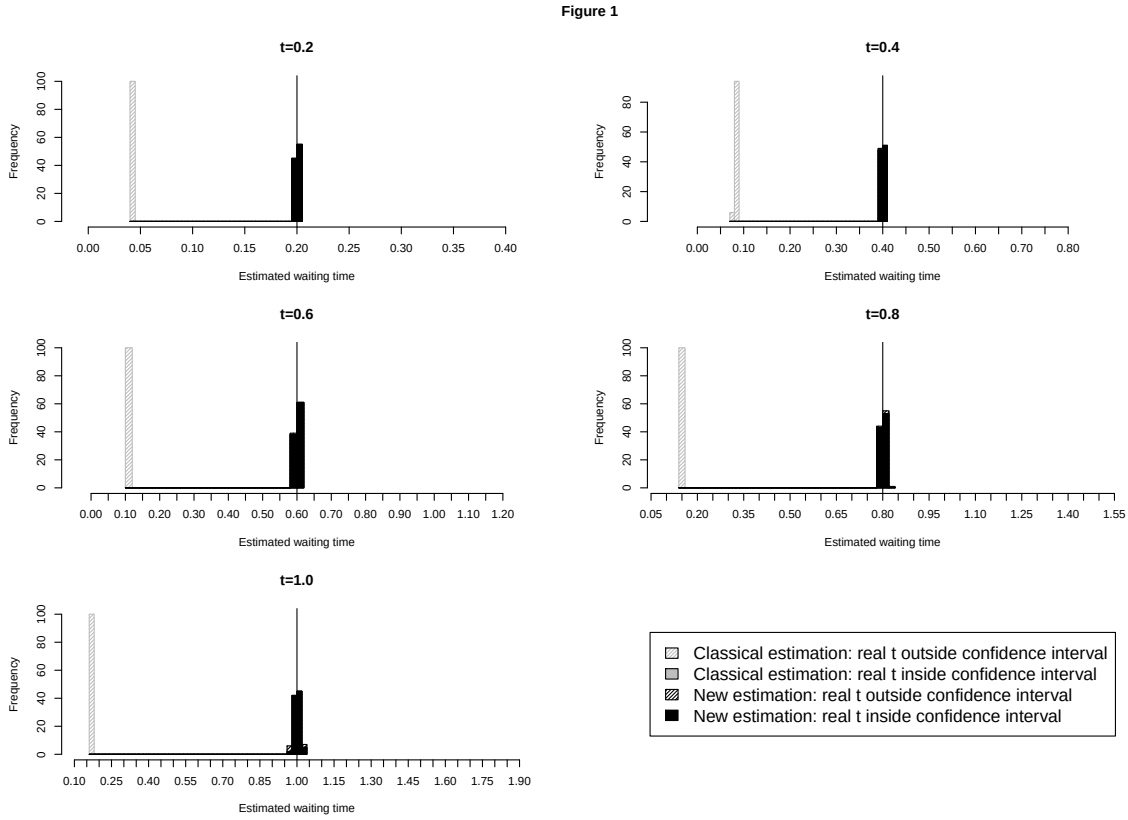

Figure 1: **Histogram of the divergence time estimators for the simulated data.** The simulation experiment has been carried out 100 times on a stationary random sequence of length $N = 1,000,000$ bp with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$ until time t, $t = 0.1, 0.2, \ldots, 1.0$. The divergence time t between the ancestral and simulated sequence was estimated by $(T^N_{\mathrm{Tamura}})_{\mathrm{est}}$ (classical estimation) and by $(T^N_C)_{\mathrm{est}}$ (new estimation). For each estimation, we checked whether the real t lies in the 95% asymptotic confidence interval.

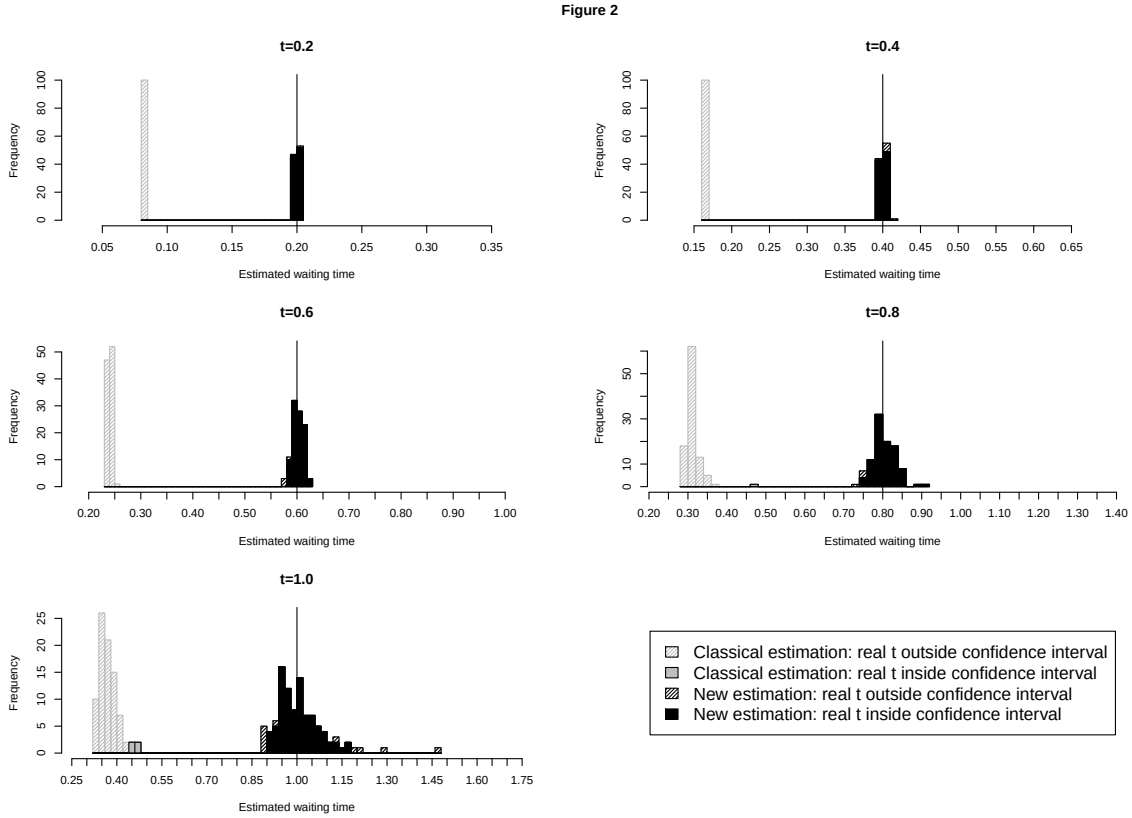

Figure 2: **Histogram of the divergence time estimators for the simulated homologous data.** The simulation experiment has been carried out 100 times on a stationary random sequence of length $N = 1,000,000$ bp with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$ until time t, $t = 0.1, 0.2, \ldots, 1.0$, resulting in two homologous sequences per simulation. The divergence time t between the two simulated homologous sequences was estimated by $[T^N_{\text{Tamura}}]_{\text{est}}$ (classical estimation) and by $[T^N_C]_{\text{est}}$ (new estimation). For each estimation, we checked whether the real t lies in the 95% asymptotic confidence interval or not.

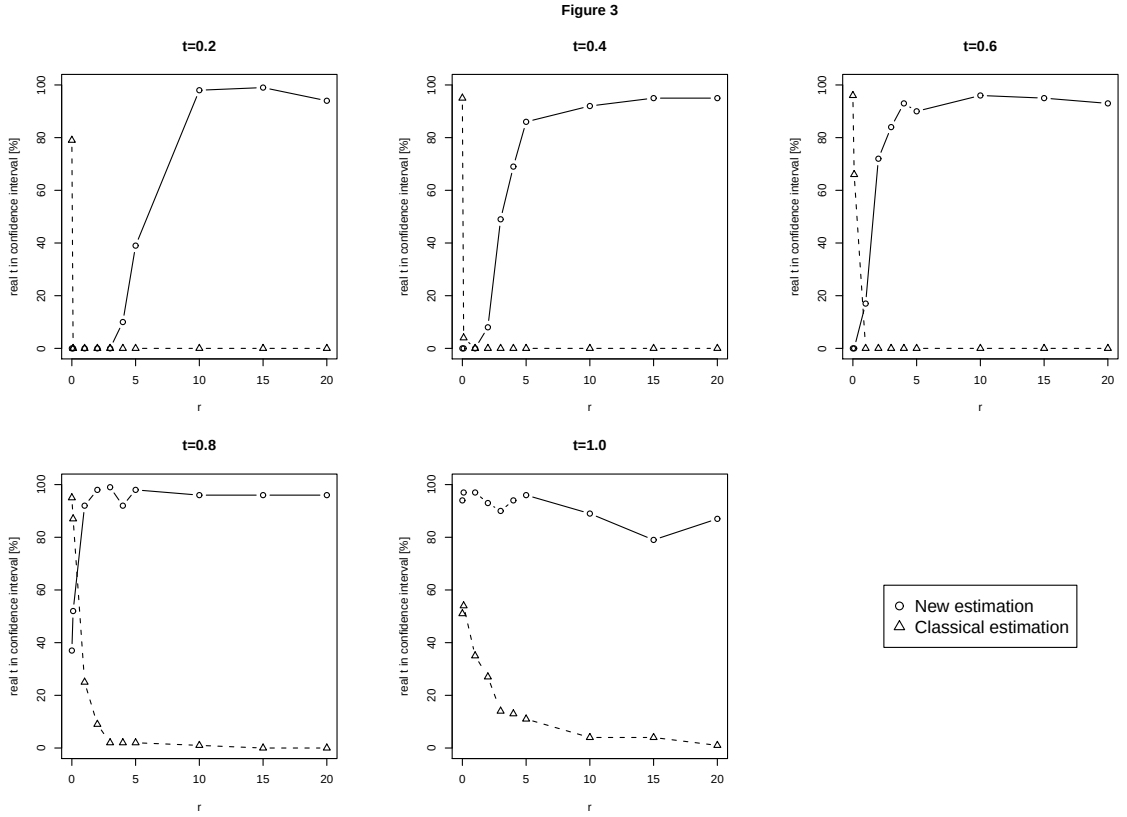

Figure 3: **Quality of the divergence time estimators.** We simulated the evolution of a stationary random sequence of length $N = 1,000,000$ bp with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and with various r for 100 times until time t, $t = 0.2, 0.4, \ldots, 1.0$, resulting in two homologous sequences per simulation. In dependence of the parameter r, we then counted how many times the real divergence time t between the two simulated homologous sequences lies in the 95%-confidence interval estimated by $[T^N_{\text{Tamura}}]_{\text{est}}$ (classical estimation) and by $[T^N_C]_{\text{est}}$ (new estimation). The count values are also given in Table 4.

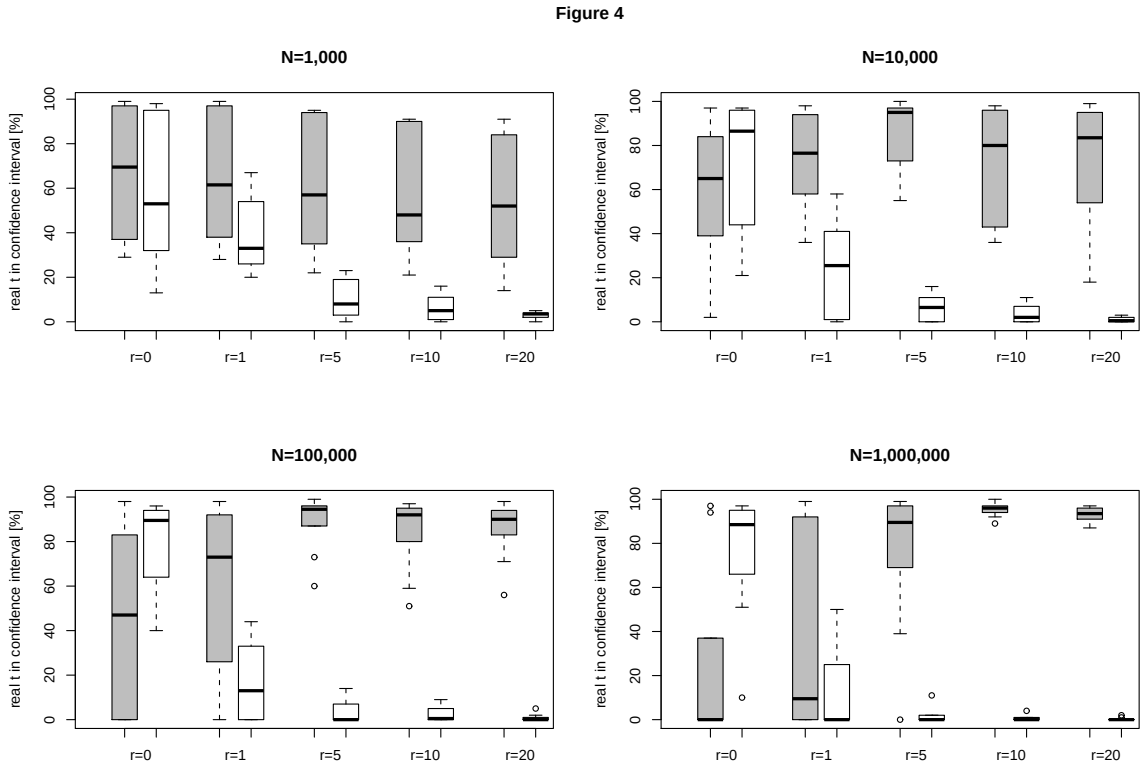

Figure 4: **Boxplot of the number of times the real t lies in the 95%-confidence interval.** For different sequence lengths N, N ranging from 10^3 to 10^6, we simulated the evolution of a stationary random sequence of length N bp with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and with various r until time t, $t = 0.1, 0.2, \ldots, 1.0$ for 100 times ending up with with two homologous sequences per simulation. In dependence of the sequence length N and the parameter r, we then counted how many times the real t lies in the 95%-confidence interval estimated by $[T^N_{\text{Tamura}}]_{\text{est}}$ (classical estimation) and by $[T^N_C]_{\text{est}}$ (new estimation). The count values for the different ts are summarized in this boxplot; color code: gray - new estimation procedure, white - classical estimation procedure.

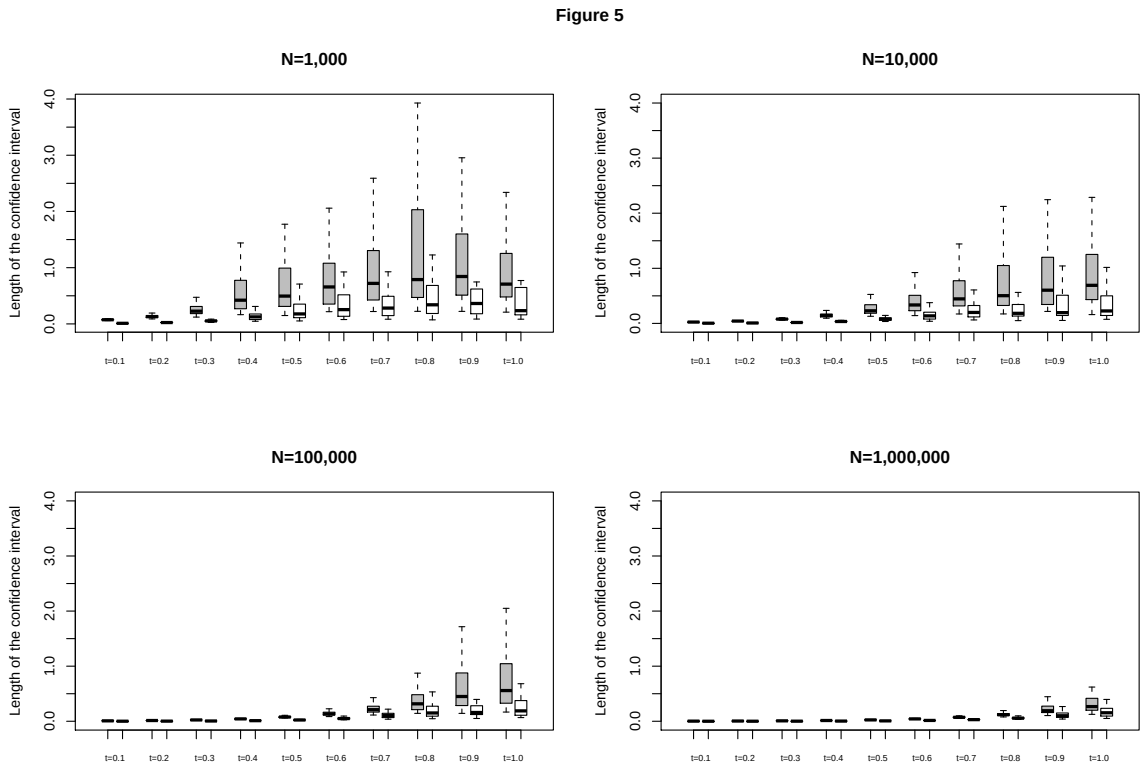

Figure 5: **Boxplot of the lengths of the 95% asymptotic confidence intervals.** Based on the results also depicted in Figure 4, for $r = 10$ we calculated the lengths of the 95% asymptotic confidence intervals estimated by the classical and by our new estimation method in dependence of the sequence length N and the real divergence time t between the two homologous sequences. Outliers are not depicted; color code: gray - new estimation procedure, white - classical estimation procedure.

Figure 6

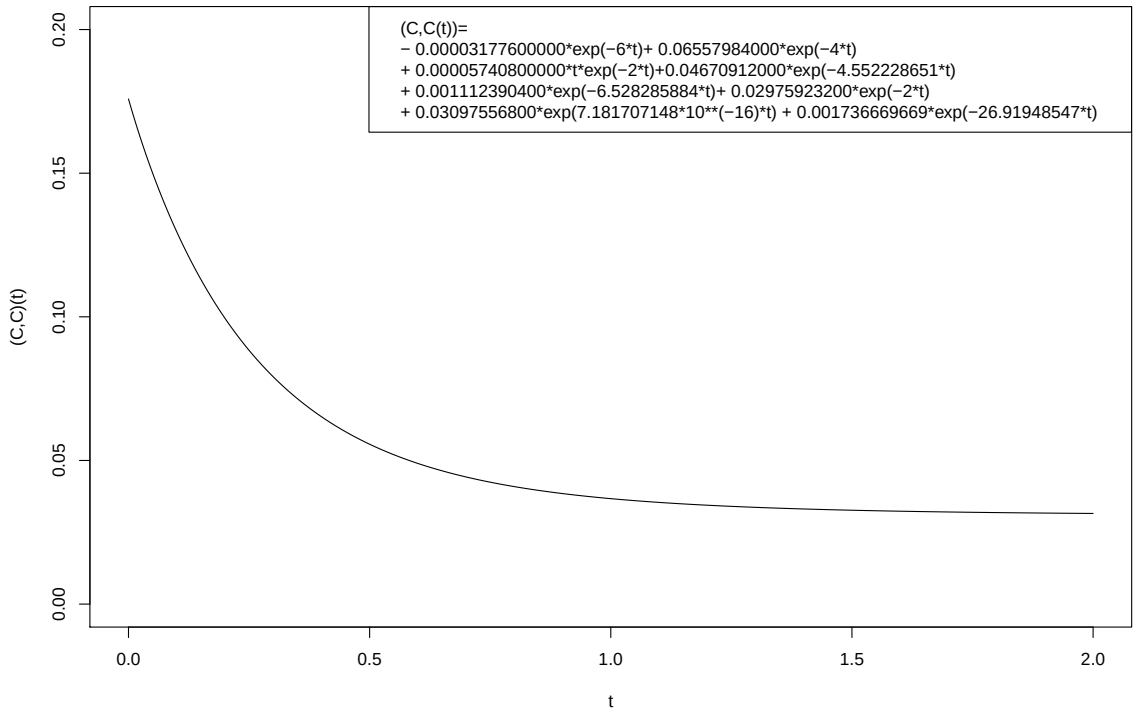

Figure 6: **The function** $(C, C)(t)$**.** Depicted is the special case of $(C, C)(t)$ with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and $r = 10$.

List of Tables

label	rate	Action: The nucleotide at site i moves ...
$\mathscr{U}_i^A$	$(1-\theta)v$	...unconditionally to A.
$\mathscr{U}_i^T$	$(1-\theta)v$	...unconditionally to T.
$\mathscr{U}_i^C$	θv	...unconditionally to C.
$\mathscr{U}_i^G$	θv	...unconditionally to G.
$\mathscr{W}_i^A$	$(1-\theta)(w-v)$	...to A provided that it is a G.
$\mathscr{W}_i^T$	$(1-\theta)(w-v)$	...to T provided that it is a C.
$\mathscr{W}_i^C$	$\theta(w-v)$	...to C provided that it is a T.
$\mathscr{W}_i^G$	$\theta(w-v)$	...to G provided that it is a A.
$\mathscr{R}_i^T$	r	...to T provided that it is a C and its right neighbor a G.
$\mathscr{R}_i^A$	r	...to A provided that it is a G and its left neighbor a C.

Table 1: **Labels, rates and actions of the Poisson processes of the T92+CpG model.** Depicted here are the labels, rates and actions of the Poisson processes necessary to simulate the evolution of site i in a circular DNA sequence under the T92+CpG model.

	RY	RA	RG	TY	TA	TG	CY	CA	CG
RY	$\cdot$	$(1-\theta)v$	θv	$(1-\theta)v$	0	0	θv	0	0
RA	v	$\cdot$	θw	0	$(1-\theta)v$	0	0	θv	0
RG	v	$(1-\theta)w$	$\cdot$	0	0	$(1-\theta)v$	0	0	θv
TY	v	0	0	$\cdot$	$(1-\theta)v$	θv	θw	0	0
TA	0	v	0	v	$\cdot$	θw	0	θw	0
TG	0	0	v	v	$(1-\theta)w$	$\cdot$	0	0	θw
CY	v	0	0	$(1-\theta)w$	0	0	$\cdot$	$(1-\theta)v$	θv
CA	0	v	0	0	$(1-\theta)w$	0	v	$\cdot$	θw
CG	0	0	v	0	0	$(1-\theta)w+r$	v	$(1-\theta)w+r$	$\cdot$

Table 2: **Infinitesimal generator $\mathbf{Q}_{\text{anc}}$.** Depicted is the infinitesimal generator Q_{anc} which rules the evolution of dinucleotides encoded in the alphabet $\mathscr{B}$ for the T92+CpG model based on Falconnet (2010a). Every diagonal term is such that the sum over the terms of its corresponding line is equal to zero.

Algorithm to solve equation $(C,C)(t) = (C,C)^N_{\text{obs}}$	
Initialization.	Set $t = 0$ and $U = U_0$.
Step 1.	If $U(CY) + U(CA) + U(CG) \leqslant (C,C)^N_{\text{obs}}$, stop and return t.
Step2.	Else, compute a new vector U according to the RK4 method, a new time $t + h$ and return to step 1.

Table 3: **Algorithm to solve equation $(\mathbf{C},\mathbf{C})(\mathbf{t}) = (\mathbf{C},\mathbf{C})^{\mathbf{N}}_{\text{obs}}$.** Based on the Runge Kutta method, we provide a numerical procedure to solve equation $(C,C)(t) = (C,C)^N_{\text{obs}}$.

Classical estimation										
r \ t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	10	79	87	95	97	96	90	95	66	51
0.1	0	0	0	4	45	66	86	87	79	54
1	0	0	0	0	0	0	3	25	50	35
2	0	0	0	0	0	0	0	9	19	27
3	0	0	0	0	0	0	0	2	15	14
4	0	0	0	0	0	0	0	2	8	13
5	0	0	0	0	0	0	0	2	2	11
10	0	0	0	0	0	0	0	1	1	4
15	0	0	0	0	0	0	0	0	2	4
20	0	0	0	0	0	0	0	0	2	1

New estimation										
r \ t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0	0	0	0	0	0	3	37	97	94
0.1	0	0	0	0	0	0	8	52	95	97
1	0	0	0	0	2	17	63	92	99	97
2	0	0	0	8	35	72	89	98	98	93
3	0	0	11	49	69	84	94	99	95	90
4	0	10	46	69	82	93	96	92	99	94
5	0	39	69	86	89	90	97	98	99	96
10	100	98	95	92	94	96	97	96	97	89
15	97	99	95	95	92	95	92	96	92	79
20	97	94	96	95	91	93	92	96	89	87

Table 4: **Quality of the estimators.** We performed 100 simulations on a stationary random sequence of length $N = 1,000,000$ bp with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and with various r until time t, $t = 0.1, 0.2, \ldots, 1.0$ ending up with two homologous sequences per simulation. In dependence of the parameter r, we counted how many times the real divergence time t between the two sequences lies in the 95%-confidence interval estimated by $[T^N_{\text{Tamura}}]_{\text{est}}$ (classical estimation) and by $[T^N_C]_{\text{est}}$ (new estimation).

Classical estimation										
r \ t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.099	0.199	0.299	0.398	0.498	0.592	0.682	0.833	0.907	0.893
0.1	0.098	0.196	0.295	0.395	0.492	0.589	0.675	0.786	0.803	0.872
1	0.087	0.174	0.262	0.349	0.439	0.518	0.612	0.662	0.699	0.727
2	0.077	0.154	0.231	0.305	0.375	0.444	0.504	0.55	0.619	0.755
3	0.07	0.14	0.209	0.276	0.347	0.43	0.477	0.511	0.541	0.609
4	0.064	0.127	0.189	0.249	0.309	0.359	0.404	0.447	0.533	0.546
5	0.059	0.116	0.173	0.229	0.286	0.335	0.389	0.465	0.513	0.5
10	0.042	0.083	0.124	0.162	0.2	0.243	0.277	0.301	0.344	0.372
15	0.033	0.064	0.095	0.127	0.158	0.189	0.224	0.256	0.272	0.284
20	0.027	0.053	0.078	0.103	0.127	0.148	0.169	0.199	0.228	0.218

New estimation										
r \ t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.129	0.234	0.344	0.452	0.559	0.647	0.743	0.886	0.936	1.055
0.1	0.126	0.23	0.336	0.439	0.547	0.66	0.76	0.917	1.025	1.158
1	0.116	0.217	0.322	0.424	0.526	0.627	0.703	0.796	0.861	0.924
2	0.109	0.209	0.312	0.411	0.506	0.6	0.677	0.747	0.8	0.994
3	0.106	0.206	0.307	0.411	0.511	0.609	0.704	0.81	0.919	1.011
4	0.104	0.204	0.305	0.406	0.507	0.599	0.678	0.74	0.883	0.952
5	0.103	0.202	0.303	0.405	0.511	0.597	0.686	0.781	0.919	0.966
10	0.1	0.199	0.3	0.397	0.498	0.613	0.732	0.802	0.973	1.027
15	0.099	0.2	0.299	0.399	0.498	0.609	0.707	0.762	0.884	0.966
20	0.099	0.2	0.298	0.394	0.499	0.587	0.68	0.802	0.953	0.92

Table 5: **Mean estimated divergence times.** Based on 100 simulations with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and with various r on a stationary random sequence of length $N = 1,000,000$ bp (homologous case), we estimated the divergence time t between the two sequences by $[T^N_{\text{Tamura}}]_{\text{est}}$ (classical estimation) and by $[T^N_C]_{\text{est}}$ (new estimation). This table provides the estimations for t averaged over the 100 simulations.

Classical estimation										
r \ t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.001	0.002	0.004	0.008	0.017	0.034	0.067	0.247	0.428	0.293
0.1	0.001	0.002	0.004	0.008	0.017	0.035	0.068	0.173	0.163	0.264
1	0.001	0.002	0.003	0.007	0.016	0.033	0.078	0.114	0.138	0.152
2	0.001	0.001	0.003	0.006	0.012	0.024	0.043	0.064	0.128	0.646
3	0.001	0.001	0.003	0.006	0.013	0.035	0.056	0.075	0.093	0.201
4	0.001	0.001	0.002	0.005	0.011	0.019	0.032	0.052	0.17	0.171
5	0.0	0.001	0.002	0.005	0.01	0.02	0.041	0.131	0.243	0.158
10	0.0	0.001	0.002	0.003	0.007	0.016	0.03	0.044	0.104	0.165
15	0.0	0.001	0.001	0.003	0.006	0.013	0.032	0.071	0.097	0.109
20	0.0	0.0	0.001	0.002	0.004	0.008	0.015	0.04	0.103	0.054

New estimation										
r \ t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.003	0.005	0.008	0.015	0.028	0.043	0.069	0.155	0.181	0.314
0.1	0.003	0.005	0.008	0.015	0.026	0.047	0.077	0.207	0.293	0.638
1	0.003	0.004	0.008	0.015	0.026	0.043	0.061	0.097	0.112	0.145
2	0.003	0.004	0.008	0.014	0.024	0.039	0.055	0.074	0.091	0.281
3	0.002	0.004	0.008	0.014	0.025	0.044	0.068	0.106	0.195	0.32
4	0.002	0.004	0.008	0.014	0.024	0.04	0.056	0.071	0.142	0.182
5	0.002	0.004	0.008	0.014	0.026	0.042	0.066	0.096	0.225	0.256
10	0.002	0.004	0.008	0.014	0.025	0.051	0.084	0.116	0.382	0.246
15	0.002	0.004	0.008	0.013	0.024	0.046	0.076	0.083	0.144	0.216
20	0.002	0.004	0.008	0.013	0.024	0.039	0.066	0.113	0.462	0.231

Table 6: **Mean lengths of the 95% asymptotic confidence intervals.** After 100 simulations with parameters $v = 1$, $w = 3$, $\theta = 0.4$ and with various r on a stationary random sequence of length $N = 1,000,000$ bp (homologous case), we estimated the divergence time t between the two sequences by $[T^N_{\text{Tamura}}]_{\text{est}}$ (classical estimation), by $[T^N_C]_{\text{est}}$ (new estimation) and the corresponding 95% asymptotic confidence intervals. This table provides the lengths of the 95%-confidence interval averaged over the 100 simulations.