BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 6 2006, pages 708–715
doi:10.1093/bioinformatics/btk001

Phylogenetics

A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model

Daniele Catanzaro[1], Raffaele Pesenti[2] and Michel C. Milinkovitch[1,*]

[1]Laboratory of Evolutionary Genetics, Institute for Molecular Biology and Medicine (IBMM), Université Libre de Bruxelles, CP300, Rue Jeener et Brachet 12, B-6041, Gosselies, Belgium and [2]DINFO, Dipartimento di Ingegneria Informatica, University of Palermo, Viale delle Scienze I-90128 Palermo, Italy

Received on November 4, 2005; revised on November 29, 2005; accepted on December 9, 2005
Advance Access publication January 5, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: The general-time-reversible (GTR) model is one of the most popular models of nucleotide substitution because it constitutes a good trade-off between mathematical tractability and biological reality. However, when it is applied for inferring evolutionary distances and/or instantaneous rate matrices, the GTR model seems more prone to inapplicability than more restrictive time-reversible models. Although it has been previously noted that the causes for intractability are caused by the impossibility of computing the logarithm of a matrix characterised by negative eigenvalues, the issue has not been investigated further.

Results: Here, we formally characterize the mathematical conditions, and discuss their biological interpretation, which lead to the inapplicability of the GTR model. We investigate the relations between, on one hand, the occurrence of negative eigenvalues and, on the other hand, both sequence length and sequence divergence. We then propose a possible re-formulation of previous procedures in terms of a non-linear optimization problem. We analytically investigate the effect of our approach on the estimated evolutionary distances and transition probability matrix. Finally, we provide an analysis on the goodness of the solution we propose. A numerical example is discussed.

Contact: mcmilink@ulb.ac.be

1 INTRODUCTION

Currently, the GTR model of DNA sequence evolution (Barry and Hartigan, 1987; Felsenstein, 1984; Lanave *et al*., 1984; Lio and Goldman, 1998; Rodriguez *et al*., 1990; Tavare, 1987; Zharkikh, 1994) is probably one of the best available trade-off between mathematical tractability and biological reality (Felsenstein, 2004, pp. 210–211; Li, 1997, pp. 81–86; Page and Holmes, 1998, pp. 152–156; Swofford *et al*., 1996, pp. 433–434; Yang, 1994. The GTR model describes DNA sequence evolution in terms of transition probabilities, $p_{ij}(t)$, from one nucleotide to another, and assumes that instantaneous substitution rate matrix, $\mathbf{R}$, remains constant over time. This stationary homogeneous Markov process can be expressed in a matrix form using Kolmogorov differential equation (Lio and Goldman, 1998):

$$\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{R}, \tag{1}$$

where $\mathbf{P}(t) = \{p_{ij}(t)\}$ is usually referred as the transition probability matrix and $\mathbf{R} = \{r_{ij}\}$ as the instantanueous substitution rate matrix (Felsenstein, 2004; Lanave *et al*., 1984; Lio and Goldman, 1998; Yang, 1994). The solution of Equation (1) is the following exponential matrix:

$$\mathbf{P}(t) = \mathbf{e}^{\mathbf{R}t}. \tag{2}$$

$\mathbf{R}$ is a real matrix with four non-positive eigenvalues [of which one is equal to zero (see, e.g. Lanave *et al*., 1984)], non-diagonal elements that must be non-negative and diagonal elements that must be the opposite of the sum of the non-diagonal elements (from the corresponding row). In turn, these conditions, together with Equation (2), imply that, for any value of t, $\mathbf{P}(t)$ is a real positive matrix characterised by four positive eigenvalues (of which one is equal to 1).

The GTR model also assumes reversibility: the net rate from nucleotides j to nucleotide i is equal to the net rate from i to j (Yang, 1994), i.e.

$$\pi_i r_{ij} = \pi_j r_{ji}. \tag{3}$$

From (2) and (3) it follows (Rodriguez *et al*., 1990)

$$\mathbf{\Pi P}(t) = \mathbf{P}(t)^{\mathrm{T}}\mathbf{\Pi}, \tag{4}$$

where $\mathbf{\Pi}$ is the diagonal matrix whose elements are the respective nucleotides equilibrium frequencies. Equation (4) can be rewritten (e.g. when considering a pair of aligned sequences separated by a time $\hat{t}$), as

$$\mathbf{P}(\hat{t}) = \mathbf{P} = \mathbf{\Pi}^{-1}(\mathbf{P}(\hat{t})^{\mathrm{T}}\mathbf{\Pi}) = \mathbf{\Pi}^{-1}\mathbf{F}^{\#}. \tag{5}$$

$\mathbf{F}^{\#}$ is called the symmetrized form (Waddell and Steel, 1997) of the divergence matrix (Rodriguez *et al*., 1990) of the observed pair of sequences. Estimating $\hat{t}$ and/or $\mathbf{R}$ [e.g. to compute $\mathbf{P}(\hat{t})$] from aligned sequences is at the core of many methods used for phylogeny inference [e.g. maximum likelihood, distance matrix methods, invariants (Felsenstein, 2004)].

*To whom correspondence should be addressed.

On the basis of conditions (1)–(5), Rodriguez *et al.* (1990) showed that the evolutionary distance $\hat{t}$ between two aligned sequences and the corresponding instantaneous substitution rate matrix **R** can be obtained by

$$\hat{t} = -\text{trace}[\boldsymbol{\Pi}\,\mathbf{log}(\mathbf{P})] \tag{6}$$

$$\mathbf{R} = \frac{\mathbf{log}(\mathbf{P})}{-\text{trace}[\boldsymbol{\Pi}\mathbf{log}(\mathbf{P})]} = \frac{\mathbf{log}(\boldsymbol{\Pi}^{-1}\mathbf{F}^{\#})}{-\text{trace}[\boldsymbol{\Pi}\mathbf{log}(\boldsymbol{\Pi}^{-1}\mathbf{F}^{\#})]}. \tag{7}$$

where $\mathbf{log}(\cdot)$ is the logarithmic matrix function defined for a square matrix with positive eigenvalues, and evaluated via diagonalization:

$$\mathbf{log}(\mathbf{P}) = \boldsymbol{\Omega}\log[\boldsymbol{\Lambda}]\boldsymbol{\Omega}^{-1} \tag{8}$$

where $\boldsymbol{\Omega}$ and $\boldsymbol{\Lambda}$ are, respectively, the eigenvector matrix of **P** and the diagonal matrix of the eigenvalues of **P**.

Rodriguez noted that this strategy is inapplicable in some cases: some conditions of inapplicability are related to the signs of the eigenvalues of **P**. More specifically, when at least one of the four eigenvalues of **P** is non-positive, the logarithm matrix function is not defined and computation of Equations (6) and (7) is not possible.

Here, starting from the seminal work of Lanave *et al.* (1984), Rodriguez *et al*. (1990), and Waddell and Steel (1997), (1) we formally characterize the mathematical and biological conditions under which the GTR model is not applicable; (2) we present sufficient criteria to a priori reject incongruent estimations of **P**; (3) we extend the estimation procedures proposed by Rodriguez *et al.* (1990) and Waddell and Steel (1997) in the form of a non-linear optimization problem that makes the GTR model always applicable; (4) we discuss the properties and the goodness of solutions obtained with our approach; (5) we suggest a procedure, different from the one proposed by Waddell and Steel (1997), to overcome the inapplicability of distance matrix methods when the entries of the distance matrix are undefined and (6) we provide a numerical example to clarify the procedure we propose.

2 APPROACH

Here, we characterize, from both a mathematical and a biological point of view, the GTR model and identify necessary and sufficient conditions that **P** must satisfy to allow the estimation of **R** using Equation (7).

2.1 On mathematical assumptions of the GTR model

As indicated earlier, the GTR model assumes that **R** is a constant matrix with a non-positive spectrum (i.e. all eigenvalues must be non-positive). In addition, let us note that, even if we relax the assumption of a constant instantaneous rate matrix **R**, the corresponding net transition matrix **P**(t) must still be characterized by positive eigenvalues. Indeed, if (1) we decompose **R** as the product of $\boldsymbol{\Pi}$ and **B**, where **B** is the symmetric matrix of rates [see (Felsenstein, 2004), p. 207] and (2) we allow **R** to be time variant (this translates into **B** being time variant as $\boldsymbol{\Pi}$ is assumed time invariant), we have

$$\mathbf{P}(t) = \mathbf{e}^{\int_0^t \mathbf{R}(\tau)\ \mathrm{d}\tau} = \mathbf{e}^{\int_0^t \boldsymbol{\Pi}\mathbf{B}(\tau)\ \mathrm{d}\tau} = \mathbf{e}^{\bar{\mathbf{R}}t}, \tag{9}$$

Table 1. The table explicitly shows the k-conditions of the Sylvester's criterion (11)

Value of k	k-th condition of the Sylvester's criterion
$k = 1$	$f_{11} > 0$
$k = 2$	$f_{11}f_{22} > f_{12}^2$
$k = 3$	$2f_{12}f_{13}f_{23} + f_{11}f_{22}f_{33} >$ $f_{23}^2 f_{11} + f_{12}^2 f_{33} + f_{13}^2 f_{22}$
$k = 4$	$f_{11}f_{22}f_{33}f_{44} + f_{14}^2 f_{23}^2 + f_{13}^2 f_{24}^2 +$ $f_{12}^2 f_{34}^2 + 2f_{12}f_{14}f_{24}f_{33} + 2f_{13}f_{14}f_{34}f_{22} +$ $2f_{12}f_{13}f_{23}f_{44} + 2f_{23}f_{24}f_{34}f_{11} > f_{14}^2 f_{22}f_{33} +$ $f_{13}^2 f_{22}f_{44} + f_{24}^2 f_{11}f_{33} + f_{12}^2 f_{33}f_{44} +$ $f_{34}^2 f_{11}f_{22} + 2f_{13}f_{14}f_{23}f_{24} + 2f_{12}f_{14}f_{23}f_{34} +$ $2f_{12}f_{13}f_{24}f_{34} + f_{23}^2 f_{11}f_{44}$

where the average instantaneous rate matrix

$$\bar{\mathbf{R}} = \frac{\int_0^t \boldsymbol{\Pi}\mathbf{B}(\tau)\ \mathrm{d}\tau}{t} \tag{10}$$

must still be characterized by a non-positive spectrum as it is the product of a positive-definite matrix $\boldsymbol{\Pi}$ times the sum of (an infinite number of) negative semi-definite matrices $\mathbf{B}(\tau)$. Hence, for any time $\hat{t}$, if we observe that at least one eigenvalue of $\mathbf{P} = \mathbf{P}(\hat{t})$ is not positive, **P** cannot be considered as a net transition matrix of a Markovian process described by Equation (2).

If not stated otherwise, we consider throughout the present paper that the matrices **P** and $\mathbf{F}^{\#}$ are computed from the observed pair of aligned sequences using the procedure described in Waddell and Steel (1997).

2.2 On the congruency between P and the GTR model

Here, we determine the conditions under which the net transition matrix **P** is congruent with the GTR model in general, and with Equation (2) in particular.

Because $\mathbf{P} = \boldsymbol{\Pi}^{-1}\ \mathbf{F}^{\#}$ and both $\boldsymbol{\Pi}$ and $\mathbf{F}^{\#}$ are symmetric, **P** is characterized by positive eigenvalues if and only if both $\boldsymbol{\Pi}$ and $\mathbf{F}^{\#}$ have positive eigenvalues (i.e. are positive-definite). We note also that $\boldsymbol{\Pi}$ is a diagonal matrix with positive diagonal entries π_i, hence, is characterized by positive eigenvalues, if all four types of nucleotides (A,T,C,G) are observed in the two aligned DNA sequences S_α and S_β (obviously, negative entries π_i would be biologically meaningless) (Keilson, 1979). Consequently, we can conclude that **P** is characterized by positive eigenvalues, hence is congruent with (2), if and only if $\mathbf{F}^{\#}$ is characterized by positive eigenvalues (i.e. is positive-definite).

As $\mathbf{F}^{\#}$ is symmetric, it is positive definite if and only if the Sylvester's criterion is satisfied [(Brinkhuis and Tikhomirov, 2005), p. 409] i.e.

$$\det(\mathbf{F}_k^{\#}) > 0, \quad k = 1, \ldots, 4 \tag{11}$$

where $\mathbf{F}_k^{\#}$ is a minor (a submatrix) of $\mathbf{F}^{\#}$ made of its first k rows and k columns. Conditions (11) are explicitly reported in Table 1, assuming that $\mathbf{F}^{\#}$ is written as follows:

$$\mathbf{F}^{\#} = \begin{pmatrix} f_{11} & f_{12} & f_{13} & f_{14} \\ f_{12} & f_{22} & f_{23} & f_{24} \\ f_{13} & f_{23} & f_{33} & f_{34} \\ f_{14} & f_{24} & f_{34} & f_{44} \end{pmatrix}. \tag{12}$$

2.3 Biological interpretation of Sylvester's conditions

When $k = 1$ the criterion indicates that, for each state (A,T,C,G), the two aligned sequences must exhibit the same state at minimum one site. Note that this condition makes a five-state GTR models [with 'gap' as a fifth state; (McGuire *et al*., 2001)] inapplicable unless one forces any pair of sequence to always share a 'gap' state at minimum one character, a condition that is not biologically meaningful. For $k > 1$, Sylvester's conditions impose constraints on (subsets of) nucleotide transition probabilities. For example, when $k = 2$, the criterion indicates that for any pair of states 1 and 2, the geometric mean between the frequency of homologous sites with identical state 1 and the frequency of homologous sites with identical state 2 must be greater than the frequency of homologous sites exhibiting different states (1 and 2). A direct consequences is that the risk of not satisfying that condition (hence, the probability of obtaining negative eigenvalues of $\mathbf{F}^{\#}$) increases with an increasing number of observed substitutions between the two aligned sequences.

Unfortunately, given the analytical complexity of the expressions when $k \geq 3$, biological interpretation of the Sylvester's conditions is much less straightforward than in the previous cases. Roughly speaking, we interpret that Sylvester's conditions when $k \geq 3$ impose a 'relative' uniformity among the different possible nucleotide transitions. To illustrate our point, let us consider, given sequences S_α and S_β, that $k = 3$ and let us assume that $f_{23} = 0$. Then, the Sylvester's condition becomes $f_{11}f_{22}f_{33} > f_{12}^2 f_{33} + f_{13}^2 f_{22}$. If we further assume $f_{22} \geq f_{33} > 0$ (i.e. $f_{22}/f_{33} \geq 1$), then Sylvester's condition becomes $f_{11}f_{22} > f_{12}^2 + f_{13}^2 f_{22}/f_{33} > f_{12}^2 + f_{13}^2$. The latter inequality implies that if $f_{23} = 0$, then it cannot occur that both f_{12} and f_{13} reach their maximum values allowed when $k = 2$.

2.4 Sufficient conditions to exclude incongruent estimations of $\mathbf{F}^{\#}$

Here, we show how one can use the Sylvester's criterion to determine some conditions that are sufficient for a priori identification of estimated $\mathbf{F}^{\#}$ matrices characterized by non-positive eigenvalues. In particular, given two generic sequences S_α and S_β, we prove that, for any pair of character states (nucleotides), $\mathbf{F}^{\#}$ cannot be estimated through the procedures proposed by Waddell and Steel (1997) (because it would be characterized by non-positive eigenvalues) when the ratio between, on one hand, the number of sites exhibiting different nucleotides in S_α and S_β and, on the other hand, the length of the sequences (l_{12}), exceeds a given threshold.

Let us consider first the Sylvester's criterion for $k = 2$. Let us define

$$\chi_1 = f_{11} + f_{12} \tag{13}$$

$$\chi_2 = f_{22} + f_{12}. \tag{14}$$

Accordingly, if one estimates $\mathbf{F}^{\#}$ as in Waddell and Steel (1997), $l_{12} = \chi_1 + \chi_2$ is, trivially, a constant that corresponds to the overall number of sites with a nucleotide of type 1 or 2 in the two sequences under consideration. Then, Sylvester's criterion for $k = 2$ can be translated into

$$(\chi_1 - f_{12})(\chi_2 - f_{12}) > f_{12}^2. \tag{15}$$

From (15), we obtain

$$f_{12} > \frac{\chi_1\chi_2}{\chi_1 + \chi_2} \leq \frac{l_{12}}{4}, \tag{16}$$

where the second inequality derives from the fact that

$$\frac{l_{12}}{4} = \sup_{\chi_1,\chi_2}\left\{\frac{\chi_1\chi_2}{\chi_1 + \chi_2} : \chi_1 + \chi_2 = l_{12}\right\} \tag{17}$$

and such value is obtained when

$$\chi_1 = \chi_2 = \frac{l_{12}}{2} \tag{18}$$

Hence, we can claim that $\mathbf{F}^{\#}$, as estimated using procedure from (Waddell and Steel, 1997), is surely not positive definite when $2f_{12} \geq (l_{12}/2)$ that, given (13) and (14) implies

$$2f_{12} \geq f_{11} + f_{22}. \tag{19}$$

In plain words, condition (19) indicates that, when two aligned sequences S_α and S_β are characterized by a number of homologous sites with different states greater than the number of homologous sites with identical states, this condition is sufficient to affirm that the procedures suggested by Waddell and Steel (1997) cannot be used.

Following the same line of reasoning, we can use Sylvester's third and fourth conditions to prove that an $\mathbf{F}^{\#}$ estimated following Waddell and Steel (1997) is surely not positive definite when at least one of the following conditions is met:

$$2(f_{12} + f_{13} + f_{23}) \geq 2(f_{11} + f_{22} + f_{33}) \tag{20}$$

$$2(f_{12} + f_{13} + f_{14} + f_{23} + f_{24} + f_{34}) \geq 3(f_{11} + f_{22} + f_{33} + f_{44}) \tag{21}$$

Note that conditions (20) and (21) are not worth checking as they can be directly derived from condition (19) by summation.

2.5 Some notes about the Logdet distance

Similarly to GTR-corrected distances, some conditions can lead to the inapplicability of the logdet correction (Lake, 1994; Lockahart *et al*., 1994; Steel, 1994). By referring to the original logdet distance formulation given by Steel (1994), we interpret that the logdet correction becomes uncomputable when the determinant of the matrix $\mathbf{F}$ is negative (its natural logarithm would be undefined). Unfortunately, the Sylvester's criterion cannot be applied in this case because $\mathbf{F}$ is not symmetric. Hence, we cannot readily extend to the logdet distance the analysis described above for GTR corrections. This issue is out of the scope of the present work and warrants additional analysis.

3 METHODS

Here, we generalize Rodriguez *et al*.'s (1990) procedure [extended in Waddell and Steel (1997)] as a non-linear optimization problem such that it returns estimates of $\mathbf{F}^{\#}$ and $\mathbf{P}$ that (1) are congruent with Equation (2), and (2) optimize a measure relevant to the evolution of the sequence S_α and S_β. Our procedure returns the same results as Rodriguez *et al*.'s (1990) when the latter yields $\mathbf{F}^{\#}$ and $\mathbf{P}$ matrices characterized by a positive spectrum.

3.1 An alternative way to estimate P

Let us consider the generic aligned homologous sequences S_α and S_β, each of length L. Let us define Γ as the set of the four different bases (A,C,G,T) and $n_{ij}^{\alpha\beta}$ as the number of nucleotides of state j of S_α that underwent substitution into state i of S_β. Analogously, let us define $n_{ij}^{\beta\alpha}$.

We then estimate matrices $\mathbf{P}$ and $\mathbf{F}^{\#}$ by determining the net transition probabilities that maximize the probability $d_{\mathbf{P}}(S_\alpha, S_\beta)$ of observing the two sequences, given $\mathbf{P}$.

$$d_{\mathbf{P}}(S_\alpha, S_\beta) = \prod_{ij} p_{ij}^{n_{ij}^{\alpha\beta}+n_{ij}^{\beta\alpha}} = \prod_{ij} (\pi_j^{-1} f_{ij})^{n_{ij}^{\alpha\beta}+n_{ij}^{\beta\alpha}} \qquad (22)$$

This is equivalent to maximizing $\mathrm{Log}(d_{\mathbf{P}}(S_\alpha, S_\beta)) = \sum_{ij} (n_{ij}^{\alpha\beta} + n_{ij}^{\beta\alpha})\mathrm{Log}(P_{ij})$. We chose this measure (among many others possible) because it is also maximized when using Rodriguez *et al.*'s (1990) procedure. When the likelihood of observing the sequences pair is maximal, $d_{\mathbf{P}}(S_\alpha, S_\beta)$ is minimal and vice versa.

Consider now the following non-linear optimization problem:

$$\max \Phi = \mathrm{Log}(d_{\mathbf{P}}(S_\alpha, S_\beta)) \qquad (23)$$

$$\mathbf{P} = \Pi^{-1}\mathbf{F}^{\#} \qquad (24)$$

$$\mathbf{F}^{\#} \succeq 0 \qquad (25)$$

$$\sum_{i\in\Gamma} f_{ij} = \pi_j \quad \forall j \in \Gamma \qquad (26)$$

$$\sum_{i\in\Gamma} \pi_j = 1 \qquad (27)$$

$$f_{ij} \geq 0 \quad \forall i,j \in \Gamma \qquad (28)$$

$$\pi_j \in Q_j \quad \forall j \in \Gamma. \qquad (29)$$

Constraint (24) imposes to search for a net transition probability matrix that can be written as the product of $\mathbf{\Pi}$ and $\mathbf{F}^{\#}$: if one would only search for a positive semi-definite net transition matrix such that its rows sum to one, one could obtain (as solution of the optimization problem) a matrix $\mathbf{P}$ that respects all the constraints but that might not be written as the product of two symmetric matrices. This would cause problems when computing $\mathbf{R}$ using Rodriguez *et al*'s (1990) procedure, because $\mathbf{\Pi}$ would be unknown. Constraint (25) imposes to search for a symmetric and positive semi-definite matrix $\mathbf{F}^{\#}$, i.e. a matrix characterised by non-negative eigenvalues. It is implemented by imposing conditions (11). Constraints (26) and (27) impose the condition of normalization on the rows of the solution and on the nucleotide frequencies. Constraints (28) and (29) trivially impose that the elements f_{ij} of the solution of the problem are non-negative and that the variables π_j are included in given sets Q_j of feasible values for nucleotide equilibrium frequencies.

The solution of problem (24)–(29) depends on which values are assumed included in the set Q_j. One reasonable choice is to directly use the average of the observed frequencies of state j in the two homologous sequences. One alternative would be that the set Q_j includes all possible values between the observed frequencies of nucleotide j in each of the two homologous sequences. Using the first option (in accordance with Rodriguez *et al*'s (1990) approach), π_j can be computed using

$$\pi_j = \frac{1}{2L}\left(n_{jj}^{\alpha\beta} + n_{jj}^{\beta\alpha} + n_{ij}^{\alpha\beta} + n_{ij}^{\beta\alpha}\right). \qquad (30)$$

This choice is a reasonable trade–off between generalisation of the model and the efficiency of solving it algorithmically. Indeed, when $\mathbf{\Pi}$ is assigned, both the objective function and the set of feasible solutions are convex, such that the solution of the problem (23) is unique (Papadimitriou and Steiglitz, 1998, p. 15). Conversely, if neither $\mathbf{F}^{\#}$ nor $\mathbf{\Pi}$ are assigned, the set of feasible solutions is not convex, such that the optimal solution might not be unique anymore. Hence, when $\mathbf{\Pi}$ is assigned the optimization problem (23)–(29) becomes:

$$\max \Phi = \mathrm{Log}[d_{\mathbf{P}}(S_\alpha, S_\beta)] \qquad (31)$$

$$\mathbf{P} = \mathbf{\Pi}^{-1}\mathbf{F}^{\#} \qquad (32)$$

$$\mathbf{F}^{\#} \succeq 0 \qquad (33)$$

$$\sum_{i\in\Gamma} f_{ij} = \pi_j \quad \forall j \in \Gamma \qquad (34)$$

$$f_{ij} \leq 0 \quad \forall i,j \in \Gamma \qquad (35)$$

The above problem can be solved by applying any standard non-linear optimization technique [see, e.g. (Bertsekas, 1999)].

Note that the optimal solution $\tilde{\mathbf{P}}$ to problem (31)–(35) might lie either inside the set of feasible solutions defined by the constrains or along its boundary. In the former case, $\mathbf{F}^{\#}$ and $\tilde{\mathbf{P}}$ are each characterized by a positive spectrum, and these correspond, respectively, to the symmetrized form of the divergency matrix and to the net transition matrix obtained by applying the procedures from Rodriguez *et al.* (1990) and Waddell and Steel (1997). In the latter case (optimal solution along the boundary of the set of feasible solutions), $\mathbf{F}^{\#}$ and $\tilde{\mathbf{P}}$ have positive spectra but at least one eigenvalue is equal to 0. The presence of null eigenvalue(s) implies that $\hat{t} = -\mathrm{trace}[\prod \log(\tilde{\mathbf{P}})] \rightarrow \infty$.

In this latter case, note that, despite $\hat{t} \rightarrow \infty$, we can still estimate $\mathbf{R}$ using

$$\mathbf{R} = \lim_{\varepsilon\rightarrow 0} \frac{\mathbf{log}(\tilde{\mathbf{P}}_\varepsilon)}{-\mathrm{trace}[\mathbf{\Pi}\mathbf{log}(\tilde{\mathbf{P}}_\varepsilon)]}, \qquad (36)$$

where $\tilde{\mathbf{P}}_\varepsilon = \frac{1}{1+\varepsilon}\mathbf{V}(\mathbf{\Lambda} + \varepsilon\mathbf{I}_\varepsilon)\mathbf{V}^{-1}$with $\mathbf{I}$, $\mathbf{V}$ and $\mathbf{\Lambda}$ that, respectively are: the identical matrix; the matrix of the eigenvectors of $\tilde{\mathbf{P}}$; and the diagonal matrix of the eigenvalues of matrix $\tilde{\mathbf{P}}$, i.e., $\tilde{\mathbf{P}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$. Roughly speaking, $\tilde{\mathbf{P}}_\varepsilon$ differs from $\tilde{\mathbf{P}}$ only by changing the eigenvalues λ_i of the latter into $\frac{\lambda_{i+\varepsilon}}{1+\varepsilon}$ in the former.

Let us finally observe that, if three eigenvalues of $\tilde{\mathbf{P}}$ are equal to 0, then

$$\tilde{\mathbf{P}} = \mathbf{P}(\infty) = \lim_{t\rightarrow\infty} \mathbf{P}(t) = \mathbf{1\Pi} \qquad (37)$$

where $\mathbf{1}$ is a matrix whose elements are all equal to 1. Actually, as shown by (Lanave *et al.*, 1984), the matrix $\mathbf{R}$ can be decomposed into:

$$r_{ij} = \pi_j + \sum_{k=2}^{4} \sqrt{\frac{\pi_j}{\pi_i}} \lambda_k w_i^{(k)} w_j^{(k)} \qquad (38)$$

where $\{w_i^{(k)}\}$ and $\{w_j^{(k)}\}$ are vectors forming an orthogonal base. Then, using (2), the transition probabilities $p_{ij}(t)$ take the following form

$$p_{ij}(t) = \pi_j + \sum_{k=2}^{4} \sqrt{\frac{\pi_j}{\pi_i}} \mathrm{e}^{-\lambda_k t} w_i^{(k)} w_j^{(k)} \qquad (39)$$

By applying the spectral decomposition theorem to $\hat{\mathbf{P}}$, we obtain:

$$\tilde{p}_{ij} = \sum_{r=1}^{4} \lambda_r v_i^{(r)} u_i^{(r)} = \pi_j + \sum_{r=2}^{4} \lambda_r v_i^{(r)} u_i^{(r)} \qquad (40)$$

where $\{v_i^{(r)}\}$ and $\{u_i^{(r)}\}$ are the left and right eigenvectors. From (39) and (40) follows (37):

$$p_{ij}(t) \rightarrow \pi_j, \quad t \rightarrow \infty \qquad (41)$$

$$\tilde{p}_{ij} \rightarrow \pi_j, \quad \lambda_{r=2\ldots4} \rightarrow 0. \qquad (42)$$

Cases in which just one or two eigenvalues of $\hat{\mathbf{P}}$ are equal to zero, are particularly interesting. Indeed, if $\hat{t} = \rightarrow\infty$, all the eigenvalues different from 1 should be equal to 0. Let us observe that using our model, or Rodriguez *et al.*'s (1990) model, one estimates continuous probability values on the basis of a finite dataset made of discrete characters (roughly speaking, we count how many sites have been subjected to substitution). Hence, the distribution of potential optimal solutions $\tilde{\mathbf{P}}$ (and, consequently, of times $\hat{t}$) of the problem (31)–(35) is not continuous. Therefore, in cases with just one or two null eigenvalues, increasing the length of the sequences to infinity would generate slightly different proportions of nucleotide substitution and a different net transition matrix, close to $\tilde{\mathbf{P}}$, but with all eigenvalues different from zero (although one or two of them very small). As a consequence, if $\tilde{\mathbf{P}}$ obtained from an observed (finite) dataset presents only one or two null eigenvalues, we conclude that $\hat{t}$ should be considered a great, but not an infinite, value.

3.2 Consequences on evolutionary distances

Application of the non-linear optimization techniques described above has major consequences on evolutionary distances. The spectrum of the net transition matrix can be considered as a measure of the level of divergence between the two analyzed sequences: the difference between the greater and the smaller eigenvalues of $\mathbf{R}t$ tends to zero as the evolutionary distance between the two observed sequences becomes smaller. Conversely, the difference between the greater and the smaller eigenvalues of $\mathbf{R}t$ tends to infinite when the evolutionary distance between the two sequences increases.

Hence, the use of distance matrix methods remains restricted to cases characterized by a positive definite matrix $\mathbf{F}^{\#}$, i.e. for pairs of sequences characterized by relatively small divergences. As, from a biological point of view, undefined (i.e. infinite) evolutionary distances are meaningless, a practical solution to overcome such a problem would be to replace $\tilde{P}$ with an appropriate $\tilde{P}_{\varepsilon}$, e.g. choosing $\varepsilon = \lambda_2^5$, where λ_2 is the smaller non-null eigenvalue of $\tilde{\mathbf{P}}$. The rationale behind such a choice is that it is generally accepted (e.g. in experimental physics) what follows. Consider two exponential processes whose asymptotic values are zero (as it occurs in our case) and such that the second process decays five times as fast as the first one. By the time the first process has assumed a value that is 1/e times its initial value, the second process has practically reached its steady state, i.e. its value is less than one hundredth of its initial value.

Furthermore, this solution has the advantage of assigning different distances for different pairs of sequences initially characterized by undefined distances (i.e. the procedure we propose provides estimates of distances) and, therefore, it might generally lead to better results than when using a fixed arbitrary value (Waddell and Steel, 1997). Note however that the approach we propose (1) introduces arbitrariness in computing the evolutionary distances (it is arbitrary to use $\varepsilon = \lambda_2^5$ rather than, e.g. $\varepsilon = \lambda_2^6$) and (2) would not be applicable if also the second eigenvalue of $\mathbf{P}$ is negative (because $\tilde{\mathrm{P}}$ would be characterized by three null eigenvalues). However this case should be very rarely encountered (i.e. it corresponds to a very high divergency between the two sequences) and should be preceded by other problems such as ambiguities in alignment.

3.3 Goodness of solutions

Let us suggest that the net transition matrix $\tilde{\mathbf{P}}$ can be considered valid (reasonable) from a biological point of view if, by random sampling of a population of sequences evolving according to $\mathbf{P}(t) = \mathbf{e}^{\mathbf{R}t}$, there is a relatively high probability to pick up a pair of sequences (S_r, S_s) whose $d_{\tilde{\mathbf{P}}}(S_r, S_s)$ is greater than or equal to $d_{\tilde{\mathbf{P}}}(S_\alpha, S_\beta)$. By indicating with n_{ij}^{xy} the number of nucleotides of type i in S_x that have undergone a substitution into the type j in S_y, with N nucleotides in each sequence, such a probability $\mathcal{P}$ can be written as follows:

$$\mathcal{P} = \sum\nolimits_{(n_{AA}^{rs}, n_{AC}^{rs}, \ldots, n_{TT}^{rs}, n_{AA}^{sr}, n_{AC}^{sr}, \ldots, n_{TT}^{sr}) \in \mathcal{K}} f(\pi, n) \tag{43}$$

where

$$\mathcal{K} = \{n_{ij}^{sr}, n_{ij}^{rs} : \sum_{(i,j)} (n_{ij}^{sr} + n_{ij}^{rs}) \log p_{ij} \leq \sum_{(i,j)} (n_{ij}^{\alpha\beta} + n_{ij}^{\beta\alpha}) \log p_{ij}\}, \tag{44}$$

$$\sum_{(ij)} n_{ij}^{rs} = N, \tag{45}$$

$$\sum_{(ij)} n_{ij}^{sr} = N, \tag{46}$$

$$n_{ij}^{rs} \geq 0, \tag{47}$$

$$n_{ij}^{sr} \geq 0, \tag{48}$$

$$i, j \in \Gamma \tag{49}$$

and

$$\begin{aligned} f(p, n) \quad = \quad & [(n_{AA}^{\alpha\beta}, n_{AC}^{\alpha\beta}, \ \ldots, n_{TT}^{\alpha\beta})! \, \Pi_{(i,j)} p_{ij}^{n_{ij}^{rs}} \\ & (n_{AA}^{\beta\alpha}, n_{AC}^{\beta\alpha}, \ \ldots, n_{TT}^{\beta\alpha})! \, \Pi_{(i,j)} p_{ij}^{n_{ij}^{sr}}]. \end{aligned} \tag{50}$$

Condition (44) imposes to consider only the indices (and therefore all the corresponding pairs of sequences) such that the observed numbers of substitutions n_{ij}^{sr} and n_{ij}^{rs} are smaller than the respective $n_{ij}^{\alpha\beta}$ and $n_{ij}^{\beta\alpha}$ under the same transition probabilities p_{ij}. Conditions (45) and (46) impose that the sums of the substitutions n_{ij}^{rs} and n_{ij}^{sr} are equal to the length N. Finally, condition (47) and (48) trivially imposes that the number of substitutions cannot be negative.

The overall probability [cf. Equation (43)] can be interpreted as the goodness of the stochastic model to predict the likelihood of observing the actual sequences. From this point of view, $\mathbf{P}(\infty)$ is better than $\tilde{\mathbf{P}}$ because the former implies a greater $\mathcal{P}$ than the latter does. In fact, for $\mathbf{P}(\infty)$, condition (44) of $\mathcal{K}$ becomes

$$\sum_{(i,j)} (n_{ij}^{sr} + n_{ij}^{rs}) \log \pi_j \leq \sum_{(i,j)} (n_{ij}^{\alpha\beta} + n_{ij}^{\beta\alpha}) \log \pi_j \tag{51}$$

and

$$\begin{aligned} f(p, n) \quad = \quad & [(n_{AA}^{\alpha\beta}, n_{AC}^{\alpha\beta}, \ \ldots, n_{TT}^{\alpha\beta})! \, \Pi_{(i,j)} \pi_j^{n_{ij}^{rs}} \\ & (n_{AA}^{\beta\alpha}, n_{AC}^{\beta\alpha}, \ \ldots, n_{TT}^{\beta\alpha})! \, \Pi_{(i,j)} \pi_j^{n_{ij}^{sr}}]. \end{aligned} \tag{52}$$

Constraint (51) is weaker than constraint (44), therefore the set $\mathcal{K}$ (hence, the overall probability $\mathcal{P}$) is greater under $\mathbf{P}(\infty)$ than under $\tilde{\mathbf{P}}$.

Notice that if one aims to obtain the transition probability matrix that maximizes $\mathcal{P}$, then one should consider the matrix $\hat{\mathbf{P}}$, i.e. the matrix characterized by all events being equiprobable (e.g. all entries = 1/4). In this situation, the set $\mathcal{K}$ becomes

$$\mathcal{K} = \{n_{ij}^{sr}, n_{ij}^{rs} : \sum_{(ij)} n_{ij}^{sr} = N, \tag{53}$$

$$\sum_{(ij)} n_{ij}^{rs} = N, \tag{54}$$

$$n_{ij}^{rs} \geq 0, \tag{55}$$

$$n_{ij}^{sr} \geq 0, \tag{56}$$

$$\forall i, j \in \Gamma. \tag{57}$$

Condition (44) in this case becomes

$$\sum_{(i,j)} (n_{ij}^{sr} + n_{ij}^{rs}) \leq \sum_{(i,j)} (n_{ij}^{\alpha\beta} + n_{ij}^{\beta\alpha}) \tag{58}$$

Constraint (58) is weaker than constraints (51), (48) and (54). Hence, when $\hat{\mathbf{P}}$ is assumed, $\mathcal{P} = 1$; i.e. the probability to obtain the observed pair of sequences is maximum. However, one should realize that, when using $\hat{\mathbf{P}}$,

$\mathcal{P}$ is always equal to 1 for any (observed or not observed) pair of sequences, i.e. all sequences are equiprobable. Our interpretation for this phenomenon is that the $\hat{\mathbf{P}}$ model does not tell us anything about the substitution process. Therefore, we suggest that, although the goodness of $\hat{\mathbf{P}}$ and of $\mathbf{P}(\infty)$ are higher than that of $\tilde{\mathbf{P}}$, the information content (regarding the substitution process of the specific pair of observed sequences) of $\tilde{\mathbf{P}}$ is higher (depending on its rank) than those of $\hat{\mathbf{P}}$ and $\mathbf{P}(\infty)$.

In summary, $\hat{\mathbf{P}}$ represents the asymptotic value of the transition probability matrix relative to any pair of sequences; $\mathbf{P}(\infty)$ represents the asymptotic value of the transition probability matrix relative to any pair of sequences characterized by a distribution $\mathbf{\Pi}$, and $\tilde{\mathbf{P}}$ is the limit transition probability matrix that best describes a substitution process [modeled by Equation (2)] for the two observed sequences characterized by a distribution $\mathbf{\Pi}$. In other words we consider that $\tilde{\mathbf{P}}$ is the most reliable matrix when the evolution of a pair of sequences is both assumed to follow (2) and characterized by negative eigenvalues of the net transition matrix $\mathbf{P}$.

4 DISCUSSION: A NUMERICAL EXAMPLE

In this section we show a numerical example of the non-linear optimization problem described above. This example is deliberately made of short and divergent sequences both to insure the occurrence of negative eigenvalues and allow the reader to easily (manually) verify each step of the procedure. Although generally not mentioned in publications, negative eigenvalues with real data do occur (e.g. the comparison ‘human versus cow’ for the cytochrome b gene third positions yield the following eigenvalues: (1., 0.423996, −0.137941, 0.111731). Whether negative eigenvalues are widespread with real data warrants further investigation.

Let us assume we want to compute the distance matrix from the following alignment:

$$\begin{bmatrix} AACGTGGCCAAAT \\ TTCGTCGTTAACC \\ CATTTCGTCACAA \\ GGTATTTCGGCCT \\ GGGACCTCGACTC \end{bmatrix}. \tag{59}$$

Let us consider the first two sequences. If we use Rodriguez *et al.*’s (1990) procedure as proposed in Waddell and Steel (1997), we obtain the following symmetrized divergency matrix

$$\mathbf{F}^{\#} = \begin{pmatrix} 0.15385 & 0.03846 & 0.0 & 0.07692 \\ 0.03846 & 0.07692 & 0.03846 & 0.11538 \\ 0.0 & 0.03846 & 0.15385 & 0.0 \\ 0.07692 & 0.11538 & 0.0 & 0.07692 \end{pmatrix}. \tag{60}$$

Accordingly, the matrix $\mathbf{\Pi}$ is

$$\mathbf{\Pi} = \begin{pmatrix} 0.269231 & 0 & 0 & 0 \\ 0 & 0.269231 & 0 & 0 \\ 0 & 0 & 0.192308 & 0 \\ 0 & 0 & 0 & 0.269231 \end{pmatrix} \tag{61}$$

with

$$\sum_{i=1}^{4} \pi_{ii} = 1.00000 \tag{62}$$

Finally, the net transition matrix $\mathbf{P}$ is

$$\mathbf{P} = \begin{pmatrix} 0.57143 & 0.14286 & 0.0 & 0.28571 \\ 0.14286 & 0.28571 & 0.14286 & 0.42857 \\ 0.0 & 0.2 & 8.0 & 0.0 \\ 0.28571 & 0.42857 & 0.0 & 0.28571 \end{pmatrix}. \tag{63}$$

This matrix is characterized by the following eigenvalues:

$$\mathbf{\Lambda}(\mathbf{P}) = \{1, 0.78886, 0.326206, -0.172209\}. \tag{64}$$

The occurrence of a negative eigenvalue makes Rodriguez *et al.*’s (1990) procedure inapplicable. Let us also observe that, in our example, all pairwise evolutionary distances are undefined, as each pairwise comparison leads to a $\mathbf{P}$ matrix characterized by negative eigenvalues. Following Waddell and Steel (1997), the distance matrix therefore is

$$\begin{bmatrix} - & \infty & \infty & \infty & \infty \\ & - & \infty & \infty & \infty \\ & & - & \infty & \infty \\ & & & - & \infty \\ & & & & - \end{bmatrix} \tag{65}$$

i.e. no distance matrix method can be used.

On the other hand, solving the non-linear optimization problem for the specific pair of sequences following the method we propose above, yields the net transition matrix

$$\tilde{\mathbf{P}} = \begin{pmatrix} 0.58030 & 0.14702 & 0 & 0.27268 \\ 0.14702 & 0.35429 & 0.16337 & 0.33531 \\ 0 & 0.22872 & 0.77128 & 0 \\ 0.27268 & 0.33531 & 0 & 0.39201 \end{pmatrix} \tag{66}$$

with eigenvalues

$$\mathbf{\Lambda}(\tilde{\mathbf{P}}) = \{1, 0.77219, 0.32569, 0\}. \tag{67}$$

Note that $\tilde{\mathbf{P}}$ satisfies the reversibility condition. The respective matrices $\mathbf{P}(\infty)$ and $\hat{\mathbf{P}}$ are

$$\mathbf{P}(\infty) = \begin{pmatrix} 0.269231 & 0.269231 & 0.192308 & 0.269231 \\ 0.269231 & 0.269231 & 0.192308 & 0.269231 \\ 0.269231 & 0.269231 & 0.192308 & 0.269231 \\ 0.269231 & 0.269231 & 0.192308 & 0.269231 \end{pmatrix} \tag{68}$$

and

$$\hat{\mathbf{P}} = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \tag{69}$$

As the evolutionary distance between the two first sequences is undefined, we propose (see Section 3.2) using $\hat{\mathbf{P}}_{\varepsilon}$, such that

$$\hat{t} = 1.84908. \tag{70}$$

By iterating the above procedure for all pairwise sequence comparisons, all entries of the distance matrix can be computed and any distance matrix method can be applied.

Let us now observe that, when computing, from the two first sequences, the substitution rate matrix $\mathbf{R}$ by using (36), we obtain

$$\mathbf{R} = \begin{pmatrix} -0.2334 & -0.2538 & 0.0574 & 0.4298 \\ -0.2538 & -1.6795 & 0.3666 & 1.5667 \\ 0.0803 & 0.5132 & -0.1602 & -0.4334 \\ 0.4298 & 1.5667 & -0.3095 & -1.6870 \end{pmatrix} \tag{71}$$

This matrix, with four negative non-diagonal elements, is not compatible with the description of a continuous Markov process and this result is not imputable to the limit (36). Indeed, an instantaneous substitution rate matrix of a Markov process must satisfy the so-called conservative hypothesis, requiring that (1) non-diagonal entries are non-negative and (2) the diagonal negative elements, row by row, are equal to the opposite of the sum of the non-diagonal elements.

The only explicit hypothesis of the estimation procedures proposed by (Rodriguez *et al*., 1990; Waddell and Steel, 1997) is reversibility as defined by Equations (4) and (5). This hypothesis is sufficient to compute $\tilde{t}$ in all cases for which $\mathbf{P}$ is congruent with (2), but is not sufficient to guarantee that the instantaneous substitution rate matrix is characterized by non-negative non-diagonal entries.

Therefore, imposing (11) when performing the non-linear optimization problem guarantees that $\tilde{\mathbf{P}}$ is congruent with (2), but does not guarantee that $\tilde{\mathbf{P}}$ a transition probability matrix of a markovian process. In such a circumstance, we propose three possible strategies.

First, one could accept $\tilde{t}$ and $\mathbf{R}$ (eventually with negative non-diagonal entries) as they are congruent with (2), but one then accepts the risk of loosing the conservative continuous markov chain hypothesis and, row per row, one uses the questionable interpretation that the eventual negative instantaneous substitution rates are decreasing substitutions in favor of the positive ones.

Second, one could accept $\tilde{t}$, and decompose the instantaneous substitution rate matrix into $\mathbf{R} = \mathbf{\Pi B}$, change the sign of all negative non-diagonal entries of $\mathbf{B}$; in this case, the conservative continuous markov chains hypothesis is guaranteed, but the optimality of $\mathbf{R}$ is lost.

Third, one could incorporate the conservative continuous markov chains hypothesis in the set of constraints of the non-linear optimization problem, i.e. we further add the constraints $\hat{t} = -\text{trace}[\Pi \log(\mathbf{P})] > 0$ and, given (7), $r_{ii} < 0$ and $r_{ij} > 0$. In this case we obtain:

$$\mathbf{R} = \begin{pmatrix} -0.767536 & 0.195231 & 0.149692 & 0.478393 \\ 0.139451 & -0.984637 & 0.19725 & 0.366612 \\ 0.149692 & 0.276149 & -0.534333 & 0.187391 \\ 0.478393 & 0.513257 & 0.187391 & -1.0324 \end{pmatrix} \tag{72}$$

and the corresponding evolutionary distance between the two first sequences is

$$\hat{t} = 1.343396. \tag{73}$$

Unfortunately such an approach is very computationally intensive and provides no guarantee that a global optimal solution can be determined.

Now, let us indicate with $d_{\mathbf{P}(\infty)}(S_\alpha, S_\beta)$ and $d_{\hat{\mathbf{P}}}(S_\alpha, S_\beta)$ the likelihood of observing the actual pair of sequences given $\mathbf{P}(\infty)$ and $\hat{\mathbf{P}}$, respectively. We obtain

$$d_{\tilde{\mathbf{P}}}(S_\alpha, S_\beta) = 3.473385 \cdot 10^{-12} \tag{74}$$

$$d_{\mathbf{P}(\infty)}(S_\alpha, S_\beta) = 7.493790 \cdot 10^{-15} \tag{75}$$

$$d_{\hat{\mathbf{P}}}(S_\alpha, S_\beta) = 3.552713 \cdot 10^{-15}. \tag{76}$$

This result confirms that $\tilde{\mathbf{P}}$ is the matrix that maximizes the likelihood of observing the specific pair of sequences. Let us compute now the overall probability to pick up a pair of sequences (S_r, S_s) that yields a likelihood greater than or equal to $d_{\tilde{\mathbf{P}}}(S_\alpha, S_\beta)$ when $\tilde{\mathbf{P}}$, $\mathbf{P}(\infty)$ or $\hat{\mathbf{P}}$ are considered. To estimate such probabilities, we generated a set of 100 sequences for each substitution matrix. We then compute the proportion of (S_α, S_β) pairs [among all pairs (ancestral sequence, generated sequence)] that exhibit a $d_{\mathbf{P}}(S_\alpha, S_\beta)$ greater than $d_{\tilde{\mathbf{P}}}(S_\alpha, S_\beta)$, $d_{\mathbf{P}(\infty)}(S_\alpha, S_\beta)$ and $d_{\hat{\mathbf{P}}}(S_\alpha, S_\beta)$. The process was repeated 100 times for each transition probability matrix $\tilde{\mathbf{P}}$, $\mathbf{P}(\infty)$ and $\hat{\mathbf{P}}$ and the three populations of sequences are characterized by the following means and variances:

$$\mathcal{P}_{\tilde{\mathbf{P}}} \simeq 1.3\% \pm 3.2 \cdot 10^{-4}\% \tag{77}$$

$$\mathcal{P}_{\mathbf{P}(\infty)} \simeq 27.4\% \pm 0.073\% \tag{78}$$

$$\mathcal{P}_{\hat{\mathbf{P}}} = 100\% \pm 0.0\% \tag{79}$$

As discussed in Section 3.3, $\mathcal{P}_{\tilde{\mathbf{P}}} \le \mathcal{P}_{\mathbf{P}(\infty)} \le \mathcal{P}_{\hat{\mathbf{P}}}$.

5 CONCLUSIONS AND PERSPECTIVES

In this paper we formally characterized the mathematical conditions that lead to the inapplicability of the published estimation procedures (Rodriguez *et al*., 1990; Waddell and Steel, 1997; Yang and Kumar, 1996) to compute evolutionary distances and instantaneous rate matrices under the GTR model of nucleotide substitution. We provided criteria to accept or reject such estimations and suggested biological interpretations of these conditions. Furthermore, we extended existing procedures (Rodriguez *et al*., 1990; Waddell and Steel, 1997, Yang and Kumar, 1996) by reformulating them in terms of a non-linear optimization problem. We analytically investigated the effect of our approach on estimated evolutionary distances, the transition probability matrix, and the instantaneous substitution rate matrix. Our formulation yields the best net transition matrix $\tilde{\mathbf{P}}$ that (1) is congruent with (2), (2) maximizes the proposed measure (22) and (3) best describes the substitution process for the specific sequence pair.

For overcoming the problem of undefined evolutionary distances, we propose a procedure that is more generally applicable and more biologically meaningful than the alternative strategy proposed by Waddell and Steel (1997).

In Section (4) we have shown that the estimation procedures proposed by Rodriguez *et al*., 1990; and by Waddell and Steel (1997) might lead in general to an estimated $\mathbf{R}$ that does not respect the conservative continuous markov chains hypothesis. This phenomenon also affects the non-linear formulation we proposed: conditions (11) are necessary and sufficient to guarantee that $\tilde{\mathbf{P}}$ is congruent with (2), but are insufficient to guarantee the conservative continuous markov chains hypothesis. We suggested three possible ways to overcome such a problem.

Finally, we observed that $\tilde{\mathbf{P}}$ has a low probability $\mathcal{P}_{\tilde{\mathbf{P}}}$. This situation seems contradictory: the matrix that maximizes the likelihood of the observed data [Equation (22)] also minimizes the probability of generating a pair of sequences with likelihood smaller than or equal to that of the observed data (43). However, we argue that a pair of sequences requiring the computation of $\tilde{\mathbf{P}}$ can not be considered as a random draw: they generate negative eigenvalues of $\mathbf{P}$. This brings us to the perspective of questioning the validity of the GTR model. First, the strength of the argument of rejecting the GTR model as a biologically valid model of nucleotide substitution depends on the frequency with which negative eigenvalues of $\mathbf{P}$ are generated with real data sequence pairs (and this points warrants further investigation). Second, there are fundamental limitations to the GTR model: it does not separate the mutation process from other factors influencing the substitution process. To investigate whether this would bring about the rejection of the GTR model and motivate the development of alternative, more complex, models, one needs to identify the biological conditions under which sets of observed sequences would be characterized by very low probabilities [Equation (43)] of being generated by the GTR model [described by Equation (2)]. Finally, another interesting issue would be to characterize the amount of time $\hat{t}$ for which, given $\tilde{\mathbf{R}}$, the probability of obtaining negative eigenvalues of $\mathbf{P}(\hat{t}) = \mathbf{e}^{\mathbf{R}\hat{t}}$ is high. The answer to this question would return the range of applicability of the GTR model. The reformulation of the GTR model as non-linear optimization problem described above is implemented in version 2 of the phylogeny inference program MetaPIGA (Lemmon and Milinkovitch, 2002), available at www.ulb.ac.be/sciences/ueg/html_files/MetaPIGA.html

ACKNOWLEDGEMENTS

We would like to thank Joseph Felsenstein and L. Gatto for helpful general discussions regarding the GTR model, as well as C. Lanave and Mike A. Steel for their invaluable comments on a previous version of this manuscript.

Conflict of Interest: none declared.

REFERENCES

Barry,D. and Hartigan,J.A. (1987) Statistical analysis of hominoid molecular evolution. *Stat. Sci.*, **2**, 191–207.

Bertsekas,D.P. (1999) *Nonlinear Programming,* 2nd edn. Hardcover.

Brinkhuis,J. and Tikhomirov,V. (2005) *Optimization: insights and applications*. Princeton University Press.

Felsenstein,J. (1984) Distance methods for inferring phylogenies: a justification. *Evolution*, **38**, 16–24.

Felsenstein,J. (2004) Inferring Phylogenies. Sinauer Associates, Sunderland, UK.

Keilson,J. (1979) *Markov Chain Models–Rarity and Exponentiality*. Springer-Verlag, New York.

Lake,J.A. (1994) Reconstructing evolutionary trees from dna and protein sequences: Paralinear distances. *Proc. Natl Acad. Sci. USA*, **91**, 1455–1459.

Lanave,C. *et al*. (1984) A new method for calculating evolutionary substitution rates. *J. Mol. Evol*., **20**, 86–93.

Lemmon,A.R. and Milinkovitch,M.C. (2002) The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proc. Natl Acad. Sci. USA*, **99**, 10516–10521.

Lio,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res*., **8**, 1233–1244.

Li,W.H. (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, UK.

Lockahart,P.J. *et al*. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol*., **11**, 605–612.

McGuire,G. (2001) Models of sequence evolution for DNA sequences containing gaps. *J. Mol. Evol*., **18**(4), 481–490.

Page,R.D.M. and Holmes,E.C. (1998) *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford, UK.

Papadimitriou,C. and Steiglitz,K. (1998) *Combinatorial Optimization, Algorithm and Complexity*. Dover Publications, Mineola, NY, USA.

Rodriguez,F. *et al*. (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol*., **142**, 485–501.

Steel,M.A. (1994) Recovering a tree from the markov the leaf colourations it generates under a markov model. *Appl. Math. Lett*., **7**, 13–23.

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) Phylogenetic inference. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *molecular systematics*. Sinauer & Associates, Sunderland, UK, pp. 407–514, chapter 11.

Tavare,S. (1987) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci*., **17**, 57–86.

Waddell,P.J. and Steel,M.A. (1997) General time reversible distances with unequal rates across sites: Mixing gamma and inverse gaussian distributions with invariant sites. *Mol. Phylogenet. Evol*., **8**, 398–414.

Yang,Z. and Kumar,S. (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rate among sites. *Mol. Biol. Evol*., **13**, 650–659.

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

Zharkikh,A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol*., **39**, 315–329.