View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UPCommons. Portal del coneixement obert de la UPC

Version dated: November 6, 2015

Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages

Jesús Fernández-Sánchez[1] and Marta Casanellas[1]

[1] *Dpt. Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Barcelona, Spain*

Corresponding author: Marta Casanellas, Dpt. Matemàtica Aplicada I
Universitat Politècnica de Catalunya, Gran Via 585, 08023-Barcelona, Spain
E-mail: marta.casanellas@upc.edu.

Abstract.— One reason why classical phylogenetic reconstruction methods fail to correctly infer the underlying topology is because they assume oversimplified models. In this paper we propose a quartet reconstruction method consistent with the most general Markov model of nucleotide substitution, which can also deal with data coming from mixtures on the same topology. Our proposed method uses phylogenetic invariants and provides a system of weights that can be used as input for quartet-based methods. We study its performance on real data and on a wide range of simulated 4-taxon data (both time-homogeneous and nonhomogeneous, with or without among-site rate heterogeneity, and with different branch length settings). We compare it to the classical methods of

neighbor-joining (with paralinear distance), maximum likelihood (with different underlying models), and maximum parsimony. Our results show that this method is accurate and robust, has a similar performance to ML when data satisfies the assumptions of both methods, and outperforms the other methods when these are based on inappropriate substitution models. If alignments are long enough, then it also outperforms other methods when some of its assumptions are violated.

[**Keywords**: phylogenetic invariants, topology reconstruction, general Markov model, heterogeneity across lineages, heterogeneity across sites, yeast]

Introduction

Usual methods of phylogenetic tree topology reconstruction are known to have limitations. For example, maximum likelihood (`ML`) is known to fail when data violate some of the underlying model assumptions (Swofford et al. 2001; Ho and Jermiin 2004; Kück et al. 2012); maximum parsimony (`MP`) is statistically inconsistent in the Felsenstein zone (Felsenstein 1978); and neighbor-joining (`NJ`) is subject to the choice of an unbiased distance and it is not as accurate as `ML` when both methods can be applied (Tateno et al. 1994). When trying to estimate relationships among distantly related organisms, neglecting heterogeneity in the substitution process across lineages or heterogeneity across sites may result in inaccurate phylogenetic estimates (see Felsenstein 1978; Fitch 1986; Yang 1994; Yang and Roberts 1995; Galtier and Gouy 1998; Ho and Jermiin 2004; Foster 2004; Kolaczkowski and Thornton 2004; Stefankovic and Vigoda 2007, among others).

Phylogenetic invariants were first introduced by Cavender and Felsenstein (1987) and Lake (1987) as a non-parametric method of phylogenetic reconstruction: they are equations satisfied by any possible joint distribution of character patterns at the leaves of a tree evolving under an evolutionary Markov model. For example, if one considers the Jukes-Cantor model on the tree 12|34, then any distribution p of patterns at the leaves satisfies the equation $p_{AACC} = p_{GGTT}$. The potential of phylogenetic invariants was the ability to deal with more general models and of detecting the topology without estimating branch lengths or substitution parameters (see Felsenstein 2004, chapter 22). In particular, they can handle heterogeneity across lineages better than other methods (Casanellas and Fernández-Sánchez 2007; Holland et al. 2013) and (some) could deal with heterogeneity across sites, as Lake's invariants did (Lake 1987). Nevertheless, only a few phylogenetic invariants were known by that time, it was not clear how to use them (Felsenstein 2004), they seemed useless for large trees, and the approach was laid aside due to the poor results obtained in simulations (Huelsenbeck 1995).

Eriksson (2005) proposed a new topology reconstruction method, `ErikSVD`, based on the work on invariants of Allman and Rhodes (2007). The underlying idea is that organizing the joint distribution p of character patterns according to a bipartition $\mathcal{A}|\mathcal{B}$ of the set of taxa gives a matrix $M_{\mathcal{A}|\mathcal{B}}(p)$ (called the *bipartition matrix* from now on, see (1) in the Methods section) of rank ≤ 4 if the bipartition is induced by an edge of the tree (and otherwise, the rank is higher). This result holds for any set of DNA sequences evolving under the most general Markov (GM) model of nucleotide substitution, also known as the Barry-Hartigan model (Barry and Hartigan 1987; Jayaswal et al. 2005; Allman and Rhodes 2008). This is the most general model assuming heterogeneity across lineages, as it allows different instantaneous rate matrices and heterogeneous composition at different parts of the tree, even locally along each branch (Jayaswal et al. 2011). Given the vector $\tilde{p}$ of relative frequencies of columns in an alignment, `ErikSVD` does not use phylogenetic invariants directly but computes the distance of the bipartition matrix $M_{\mathcal{A}|\mathcal{B}}(\tilde{p})$ to the set

of matrices of rank ≤ 4. The shorter the distance, the more likely is the bipartition $A|B$ to come from an edge of the tree underlying the data.

However, the original `ErikSVD` method turned out not to be accurate enough to compete against standard methods (Eriksson 2005), especially in the presence of long-branches and short alignments. Here we revisit `ErikSVD` by adjusting the bipartition matrix: we normalize it by column (respectively, row) sums so that we obtain the transition matrix from the states of one side of the bipartition to the other. This adjustment is made to take into account that the rank of the bipartition matrix obtained from empirical distributions could be affected by the presence of long-branch attraction situations. Indeed, in these situations the probability of observing the same nucleotide at two non-sister species a and b can be large, so the corresponding row at the bipartition matrix $M_{ab|others}(\tilde{p})$ has large entries, which can distort its theoretical rank (see Appendix 1 for an example). The original `ErikSVD` was already *statistically consistent* (that is, as the empirical distribution approaches the theoretical distribution, the probability of correctly reconstructing the tree goes to one) and so is the method proposed here (see the Methods section) which will be called `Erik+2`.

`Erik+2` is model-based as it assumes a general Markov model of evolution (and it could also be redesigned to incorporate either more restrictive Markov models or amino acid substitution models), but is non-parametric in the sense that it does not attempt to recover the parameters of the model. Moreover, the theoretical background of `Erik+2` allows it to be applied on heterogeneous data across sites evolved on the same tree topology under the GM model (Jayaswal et al. 2014): that is, a parameter m can be introduced so that `Erik+2` considers the sites of the alignment to be divided into m categories, each evolving on the same topology but with (possibly) different Markov substitution matrices and nucleotide distribution at the root –this is often called an *m-mixture* (Stefankovic and Vigoda 2007). For example, discrete-gamma rates or the heterogeneous tree in Kolaczkowski and Thornton (2004) are instances of mixtures, and `ML` is known to fail under

these conditions even when consistent underlying homogeneous models are considered (Kolaczkowski and Thornton 2004; Kück et al. 2012). For m-mixtures, the rank of the bipartition matrix induced by an edge is not larger than $4m$ (Rhodes and Sullivant 2012, e.g.) so that in this case we compute the distance to matrices of rank $\leq 4m$.

We develop `Erik+2` on 4-taxon trees and study its performance on simulated and real data. Using computer simulations we compare it to the classical methods `ML`, `NJ`, `MP` and to the original `ErikSVD` in many different scenarios. We chose quartets because they are the smallest building blocks of phylogenetic reconstruction (Ranwez and Gascuel 2001) and they are widely used as a hint of efficiency and robustness of the method under study (Huelsenbeck 1995). Another reason to focus on quartets is that the number of possible patterns increases exponentially with the number of leaves, thus producing poor estimations of the pattern distribution for large number of leaves. `Erik+2` evaluates the three possible quartet topologies and returns a system of weights that can be used as input for quartet-based methods (see the Methods section).

Some of our computer simulations are generated under the general Markov process that underlies `Erik+2` and some are based on the most general time-reversible (GTR) and homogeneous across lineages model (*homGTR* from now on). We also simulate heterogenous data across sites by generating either 2-mixtures on the same topology evolving under the GM model or gamma continuously distributed rates across sites under the homGTR model. Throughout the paper `NJ` has been considered with the paralinear distance, and `ML` computations have been based on continuous-time models (with parameters to be estimated by the method) considering homogeneity or heterogeneity across lineages and sites depending on the situation.

The performance of `Erik+2` on real data is analyzed on the eight species of yeast studied in Rokas et al. (2003) with the concatenated alignment provided by Jayaswal et al. (2014). We investigate whether the quartets output by `Erik+2`, `ErikSVD` and `ML` support the tree T of Rokas et al. (2003) or the alternative tree T' of Phillips et al. (2004), and

the mixture model proposed by Jayaswal et al. (2014).

Methods

ErikSVD and Erik+2 methods

`Erik+2` arises as a variation of the method described by Eriksson (2005) by normalizing certain matrices obtained from an alignment of nucleotide sequences. As in the original method, the information contained in the alignment is recorded as a vector $\tilde{p}$ whose coordinates are the observed relative frequencies of possible patterns at the leaves. In the case of an alignment of four taxa 1,2,3,4, each possible (trivalent) topology is determined by a bipartition of the taxa: 12|34, 13|24 or 14|23. For each bipartition $A|B$, a matrix $M_{\mathcal{A}|B}(\tilde{p})$ is considered by rearranging the coordinates of $\tilde{p}$ according to it, so that the rows of the matrix $M_{\mathcal{A}|\mathcal{B}}(\tilde{p})$ are indexed by all possible observations at the taxa in A, and similarly for columns and observations at the taxa in B. For example, the (AG, CT)-entry of $M_{12|34}(\tilde{p})$ is the relative frequency $\tilde{p}_{AGCT}$ of the pattern $AGCT$ in the alignment. The same entry in $M_{13|24}(\tilde{p})$ corresponds to the relative frequency $\tilde{p}_{ACGT}$ of $ACGT$.

$$M_{12|34}(\tilde{p}) = \begin{array}{c} \\ AA** \\ AC** \\ AG** \\ \vdots \\ TT** \end{array} \begin{array}{c} \begin{array}{ccccc} **AA & **AC & **AG & \dots & **TT \end{array} \\ \begin{pmatrix} \tilde{p}_{AAAA} & \tilde{p}_{AAAC} & \tilde{p}_{AAAG} & \cdots & \tilde{p}_{AATT} \\ \tilde{p}_{ACAA} & \tilde{p}_{ACAC} & \tilde{p}_{ACAG} & \cdots & \tilde{p}_{ACTT} \\ \tilde{p}_{AGAA} & \tilde{p}_{AGAC} & \tilde{p}_{AGAG} & \cdots & \tilde{p}_{AGTT} \\ \vdots & \vdots & \vdots & & \vdots \\ \tilde{p}_{TTAA} & \tilde{p}_{TTAC} & \tilde{p}_{TTAG} & \cdots & \tilde{p}_{TTTT} \end{pmatrix} \end{array} \tag{1}$$

Assume that the coordinates of $\tilde{p}$ are the empirical estimates of the theoretical joint distribution p at the leaves of a tree T evolving under the GM model, say $T = 12|34$. Then the key point is Theorem 19.5 of Eriksson (2005) (see also Casanellas and Fernández-Sánchez (2010)) that claims that the rank of $M_{\mathcal{A}|\mathcal{B}}(p)$ is 4 if $\mathcal{A}|\mathcal{B} = 12|34$, and 4^2 otherwise

(if the substitution matrices that generated p were general enough). Eriksson's idea is to compute the *Frobenius distance* (that is, the Euclidean distance if we view the matrices as elements in $\mathbb{R}^{4^2 \times 4^2}$, Demmel 1997) d_4 of the three matrices $M_{12|34}(\tilde{p})$, $M_{13|24}(\tilde{p})$ and $M_{14|23}(\tilde{p})$ to the space of matrices of rank ≤ 4. In this manner, one derives which of the three matrices is closer to having rank ≤ 4. Note that the Frobenius distance of a matrix M to the set of matrices of rank $\leq k$, $\mathrm{d_k}(M)$, is easily computed in terms of the singular values of M (Eckart and Young 1936).

The main motivation for the variation introduced in `Erik+2` arises from the observation that the presence of short branches may seriously affect this distance when it is computed from short alignments. For example, for a tree with small a and large b (a tree in the Felsenstein zone; Figure 1.a), the distance of $M_{13|24}(\tilde{p})$ to 4-rank matrices can be smaller than that of $M_{12|34}(\tilde{p})$ (see also Appendix 1). The reason is that a small a implies that the probability of observing the same nucleotide at leaves 2 and 4 is high, so columns in $M_{13|24}(\tilde{p})$ indexed by $*A*A$, $*C*C$, $*G*G$, or $*T*T$ capture most of the non-zero entries in the matrix, while other columns may only have few nonzero entries. Thus $M_{13|24}(\tilde{p})$ is not far from having rank 4, even if 13|24 is not the correct topology.

Moreover, if p evolves on the 12|34 topology, the sum of the entries of a column of $M_{12|34}(p)$ depend on the substitution matrices at the branches 3,4. In the spirit of producing a more robust score that is less influenced by the peripheral substitution processes (see for example the idea of using *Markov invariants* in Holland et al. (2013)), we normalize these sums. By dividing any non-zero column by the sum of its entries, all the non-zero columns become of the same weight. As the same situation may occur with rows, we also need to normalize the matrix by row sums. In this way, each matrix $M_{\mathcal{A}|\mathcal{B}}(\tilde{p})$ gives rise to a pair of *transition* matrices $M_{\mathcal{A}\rightarrow\mathcal{B}}(\tilde{p})$ and $M_{\mathcal{A}\leftarrow\mathcal{B}}(\tilde{p})$, obtained by column and row sum adjustment, respectively:

$$M_{12\to 34}(\tilde{p}) = \begin{pmatrix} \frac{\tilde{p}_{AAAA}}{\tilde{p}_{AA++}} & \frac{\tilde{p}_{AAAC}}{\tilde{p}_{AA++}} & \cdots & \frac{\tilde{p}_{AATT}}{\tilde{p}_{AA++}} \\ \frac{\tilde{p}_{ACAA}}{\tilde{p}_{AC++}} & \frac{\tilde{p}_{ACAC}}{\tilde{p}_{AC++}} & \cdots & \frac{\tilde{p}_{ACTT}}{\tilde{p}_{AC++}} \\ \frac{\tilde{p}_{AGAA}}{\tilde{p}_{AG++}} & \frac{\tilde{p}_{AGAC}}{\tilde{p}_{AG++}} & \cdots & \frac{\tilde{p}_{AGTT}}{\tilde{p}_{AG++}} \\ \cdots & \cdots & & \cdots \end{pmatrix} \quad M_{12\leftarrow 34}(\tilde{p}) = \begin{pmatrix} \frac{\tilde{p}_{AAAA}}{\tilde{p}_{++AA}} & \frac{\tilde{p}_{AAAC}}{\tilde{p}_{++AC}} & \cdots & \frac{\tilde{p}_{AATT}}{\tilde{p}_{++TT}} \\ \frac{\tilde{p}_{ACAA}}{\tilde{p}_{++AA}} & \frac{\tilde{p}_{ACAC}}{\tilde{p}_{++AC}} & \cdots & \frac{\tilde{p}_{ACTT}}{\tilde{p}_{++TT}} \\ \frac{\tilde{p}_{AGAA}}{\tilde{p}_{++AA}} & \frac{\tilde{p}_{AGAC}}{\tilde{p}_{++AC}} & \cdots & \frac{\tilde{p}_{AGTT}}{\tilde{p}_{++TT}} \\ \cdots & \cdots & & \cdots \end{pmatrix}.$$

We give a score to any tree $T_{\mathcal{A}|\mathcal{B}}$ as

$$\mathrm{sc}(T_{\mathcal{A}|\mathcal{B}}) := \frac{\mathrm{d}_4\left(M_{\mathcal{A}\to\mathcal{B}}(\tilde{p})\right) + \mathrm{d}_4\left(M_{\mathcal{A}\leftarrow\mathcal{B}}(\tilde{p})\right)}{2}.$$

`Erik+2` outputs the topology with smallest score (notice that the smaller the score, the more reliable the topology $T_{\mathcal{A}|\mathcal{B}}$ is). As the empirical distribution $\tilde{p}$ approaches the theoretical distribution p, the transition matrices $M_{\mathcal{A}\to\mathcal{B}}(\tilde{p})$ and $M_{\mathcal{B}\to\mathcal{A}}(\tilde{p})$ approach the theoretical transition matrices. These have rank 4 for the correct topology because they have the same rank as the theoretical bipartition matrices (as they are obtained from them by dividing rows/columns by scalars). Therefore $\mathrm{d}_4\left(M_{\mathcal{A}\to\mathcal{B}}(\tilde{p})\right)$ and $\mathrm{d}_4\left(M_{\mathcal{B}\to\mathcal{A}}(\tilde{p})\right)$ tend to 0 when $\tilde{p}$ approaches the theoretical distribution (as the Frobenius distance is a continuous function) and thus `Erik+2` is statistically consistent.

In order to avoid bias in the rank of the bipartition matrices due to insufficient data, in the adjustment by rows we disregard rows whose sum is not larger than $2/N$ where N is the alignment length (same for columns).

`Erik+2` also provides normalized weights that can be used as input for weighted quartet-based methods (Strimmer and von Haeseler 2006). Indeed, the score above is turned into a confidence weight by inverting it and normalizing so that the overall sum of weights is 1:

$$w(T_{\mathcal{A}|\mathcal{B}}) := \frac{\mathrm{sc}(T_{\mathcal{A}|\mathcal{B}})^{-1}}{\sum_T \mathrm{sc}(T)^{-1}}.$$

The basic model underlying `Erik+2` and `ErikSVD` assumes that all sites in the alignment are independently and identically distributed according to a general Markov model. There is no extra assumption about the shape of the substitution matrices (nor stationarity, nor

time-reversibility, nor global or local homogeneity). But in `Erik+2` we relax the i.i.d hypotheses and allow heterogeneity across sites by considering mixtures in the sense of Kolaczkowski and Thornton (2004) and Stefankovic and Vigoda (2007). That is, a single tree topology T is considered but we allow m categories of Markov processes on T defined by m sets $(\sigma_1, \ldots, \sigma_m)$ of substitution parameters and nucleotide distribution at the root. The proportion of sites contributed by the i-th tree (T, σ_i) is denoted by p_i and the joint distribution at the leaves of T follows an *m-mixture distribution*: $\sum_i p_i P(T, \sigma_i)$. A parameter $m \in \{1, 2, 3\}$ can be passed to `Erik+2` to adapt the method to consider m categories. In this case, we compute the distance $\mathrm{d_{4m}}$ to matrices of rank $\leq 4m$ (for m-mixtures the bipartition matrices can be viewed as the sum of m matrices which have rank ≤ 4 if the bipartition is an edge split, so that the bipartition matrix has rank $\leq 4m$). The restriction to up to 3 categories is only due to theoretical results about non-identifiability for quartets with four or more categories (there would be 255 parameters in a 4-mixture, which already fills the whole space of pattern distributions, see Casanellas et al. (2012)); but for up to 3 categories the quartet is known to be identifiable (Allman and Rhodes 2006).

We had also developed different modifications of the original method of Eriksson, all of them showing lower success than the version considered here. For example we have tried different options to provide a single value from the distances $\mathrm{d_4}$ of the two transition matrices: the 2-norm, the ∞-norm or even the minimum. We also tried normalizing by rows and columns at the same time (with poor results), and also considering the 2-norm instead of the Frobenius norm (similar results). Therefore in this paper we only present the results corresponding to the method `Erik+2` explained here.

Maximum likelihood

Heterogeneity across lineages.— In order to estimate a heterogeneous across lineages

model we used the software `bppml` of the Bio++ package (Dutheil and Boussau 2008). The software had to infer a GTR rate matrix per edge, the branch lengths, and the equilibrium distribution on a rooted quartet tree with either

- homogeneity across sites, `ML`(GTR) henceforth, or
- discrete gamma rates with two/three categories, `ML`(GTR+2Γ) / `ML`(GTR+3Γ), or
- two categories plus invariable sites, `ML`(GTR+2Γ+I).

Homogeneity across lineages.— When we restricted the likelihood computations to homogeneous across lineages continuous-time models, we used PAML (Yang 1997) to estimate the rate matrix Q (a unique rate matrix for the whole tree), the equilibrium distribution, and the branch lengths on a rooted quartet tree. Depending on the setting, the method infers either

- an unrestricted rate matrix Q (model UNREST in PAML documentation) leading to the most general *continuous*-time *homogeneous* Markov model, which is denoted as `ML`(homGMc), or
- a rate matrix Q restricted to the GTR model, `ML`(homGTR), with possibly auto-discrete gamma rates, `ML`(homGTR+Γ).

For the `ML` computations we waited up to 60 seconds for convergence on each tree topology and if it did not converge, we treated it as failed (because we cannot compare likelihoods in this case). It is worth pointing out that, usually, `ML` was not convergent only for the incorrect topologies.

Neighbor-joining

As far as neighbor-joining is concerned, the paralinear distance (Lake 1994) was used throughout the paper to estimate pairwise divergences in NJ. This distance is based on the GM model and has shown to be very useful for phylogenetic inference when nucleotide sequences are nonstationary (Gu and Li 1996). We also tested `NJ` with the other distances proposed in this last reference but the best results in the scenarios considered (always on homogenous across sites data) were invariantly obtained by the paralinear distance.

Description of the data

In order to test different methods, we simulated data under different scenarios:

Felsenstein zone.— We consider trees subject to long-branch attraction, also known as trees in the Felsenstein zone. To this end, on the tree of Figure 1.a we fix $a = 0.05$, $b = 0.75$, and let the internal branch length c vary in the range $[0.01, 0.4]$ (or either we fix $c = 0.05$ and call it the *Felsenstein tree*). For each set of branch lengths, we generated one hundred alignments of different lengths.

Tree space.— We adopt a similar approach to Huelsenbeck (1995) to test different methods. More precisely, we evaluate the methods on a *tree space* (see Figure 1.b) where the quartets are as in Figure 1.a with $c = a$, and the branch lengths a and b vary between 0 and 1.5 in steps of 0.02. For each pair a, b we generate one hundred alignments of a fixed length and represent the success of different methods in recovering the right topology.

Farris tree.— We also consider trees with two incident pendant edges of length equal to 0.5 and the other three edges of equal length varying from 0.01 to 0.05 (Figure S5 in Appendix 2).

Mixture data.— As mentioned above, one of the main features of `Erik+2` is that it can deal with different categories of evolutionary rates. In order to test its accuracy on such setting, we use the approach of Kolaczkowski and Thornton (2004). We consider two

categories of the same size both evolving under the GM model on the tree of Figure 1.a: the first category corresponds to branch lengths $a = 0.05$, $b = 0.75$, while the second corresponds to $a = 0.75$ and $b = 0.05$. The internal branch length takes the same value in both categories and varies from 0.01 to 0.4 in steps of 0.05.

Simulations under the GM model.— To generate data under the general Markov model, we have used `GenNon-h` (Kedzierska and Casanellas 2012). Given a set of branch lengths (understood as the expected number of substitutions per site) and the tree topology of Fig. 1 rooted at the parent node of leaves 3 and 4, this software generates a random distribution of nucleotides at the root and random substitution matrices with the expected amount of substitutions per site, and lets nucleotides evolve according to this Markov process on the tree (Casanellas and Kedzierska 2013).

Simulations under a time-homogeneous GTR model.— In order to generate data evolving under a homGTR model (with or without continuous gamma-rates) we have used `Seq-gen` (Rambaut and Grassly 1997). We used uniform equilibrium distribution, and the rate matrix underlying `Seq-gen` alignments on the tree space used in Fig. 4 and Appendix 2.S2 had rates 2 (A→C), 7 (A→G), 4 (A→T), 3 (C→G), 1 (C→T), 5 (G→T). The rate matrix underlying GTR+Γ had rates 2 (A→C), 5 (A→G), 3 (A→T), 4 (C→G), 1 (C→T), 2 (G→T) and the sites were varied according to a gamma distribution with parameter $\alpha = \beta$ in the range $[0.1, 2]$ varying in steps of 0.1. Small values of this parameter indicate a lot of variation across sites (Yang 1993).

Real data.— We considered the data provided by Jayaswal et al. (2014) with 42 337 second codon positions of 106 orthologous genes of *Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. castellii*, *S. kluyveri*, *S. bayanus*, and *Candida albicans*. The phylogenetic tree of these species was originally studied in Rokas et al. (2003), where a tree topology T was identified with 100% bootstrap support for the concatenated alignment

of these genes. This tree is widely accepted by the community but its correct inference is known to depend on the consideration of heterogeneity across lineages (Rokas et al. 2003; Phillips et al. 2004; Jayaswal et al. 2014). For example, Phillips et al. (2004) obtain an alternative tree T' with 100% bootstrap using the method of minimum evolution, but identified the incorrect handling of compositional bias as responsible for this inconsistency. Moreover, according to Jayaswal et al. (2014) these data are best modeled by taking into account heterogeneity across lineages plus two different rate categories 2Γ and invariable sites I. In our setting, this would involve a 3-mixture distribution.

Results

The source code of `Erik+2` is publicly available via Dryad at `doi:10.5061/dryad.mh850`. It is also available at the webpage http://geomap.ma1.upc.edu/links/erik2.

We present the performance of the new method `Erik+2`, `ErikSVD`, `ML`, `NJ` and `MP`, on quartet reconstruction on different simulated data.

Homogeneity across sites

First of all we evaluate `Erik+2`, `ML`, `MP` and `ErikSVD` on the Felsenstein zone on data generated under the GM model (Fig. 2). In this figure two models underlying `ML` computations have been considered: the most general homogeneous continuous-time model, `ML`(homGMc), and a (heterogeneous across lineages) GTR model, `ML`(GTR). `ML` has a similar performance with both models. We observe that `Erik+2` is more accurate than `ErikSVD` in general and especially when the interior branch length is short (only for length 1 000 and $c \in (0.13, 0.25)$ `ErikSVD` outperforms slightly `Erik+2`). Both versions of `ML` perform better than `Erik+2` for 1 000 bp., `Erik+2` and `ML` perform similarly for 10 000 sites, but `Erik+2` outperforms `ML` (in its both versions) for 100 000 sites and short internal

branch. Notice incidentally that the accuracy of MP does not seem to increase as the length of the alignment grows.

In a more complete study, the methods Erik+2, ML, and NJ have been tested on the *tree space* for data generated under the GM model (Fig. 3) and under a homogeneous GTR model (Fig. 4). Both ML(homGMc) and NJ have lower accuracy than Erik+2 (Fig. 3), as was expected under data that violates the assumptions of ML and NJ. While Erik+2 and NJ drastically increase their accuracy when the alignment length is multiplied by 10, ML does not (Fig. 4 and, very specially, Fig. 3 where the substitution model assumed by ML is incorrect). Under the homogenous GTR model and for length 10 000 (Fig. 4), Erik+2 performs slightly better thatn ML(homGTR) (although the difference does not seem significant). One possible explanation is suggested by the presence of long branches, which may distort the performance of ML even when the correct model is assumed (Kück et al. 2012). Erik+2 also outperforms ErikSVD on this tree space (Fig. 1.b and Fig. S1 in Appendix 2). We also present the average success achieved by these methods on this tree space for alignments of length 500 bp. and 1 000 bp. (Table 1).

It is worth pointing out that, for alignments of 1 000 bp. evolving under the homogeneous GTR model, ML seems to outperform Erik+2 in the Felsenstein zone when it estimates exactly this model, ML(homGTR) (Fig. 4). However, for length 10 000, Erik+2 already outperforms ML(homGTR). Moreover, the global accuracy of ML(homGMc) drastically drops when applied to data obtained under the GM model (Fig. 3). Notice also that whereas the accuracy of NJ and ML drops when all branches are long (top right corner), the performance of Erik+2 seems less sensitive to long-branches.

We have also evaluated the version of Erik+2 with 2-mixtures ($m = 2$) on the same data (Fig. S3 in Appendix 2). The accuracy obtained for alignments of 1 000 bp. is similar to that of Erik+2 with $m = 1$ (the means are 0.790 and 0.803, respectively), and hence the choice $m = 2$ appears as a good option when alignments are long enough and we ignore whether the data comes from mixtures or not (see also Fig. S4 in Appendix 2).

For completeness, we also include results on the performance of `Erik+2` and `ML` (GTR) on the Farris tree with data simulated under the GM model (Fig. S5 in Appendix 2). In this case, `ML` performs slightly better than `Erik+2`.

Heterogeneity across sites

We also generated data under the homogeneous GTR model with sites varying according to a gamma distribution on the Felsenstein tree. While this setting violates the hypotheses of the model underlying `Erik+2` and `ErikSVD`, in this case maximum likelihood is estimating a homogeneous GTR model with rates varying according to the auto-discrete gamma model `ML`(homGTR+Γ) (Yang 1994). We observe that `Erik+2` manages to overcome the violation of its hypotheses giving 100% success already for 10 000 bp., while `ErikSVD` gives terrible results for 1 000 and 10 000 bp., but excellent results for 100 000 bp. (Figure 5.a). `ML` is more successful than `Erik+2` for 1 000 bp., but both methods have a similar performance on longer alignments. On the same data we also tested `MP`, obtaining in all cases the incorrect tree 13|24 (and therefore we do not represent the corresponding 0% line in the figure).

In order to check the accuracy of `Erik+2` on data with two categories of evolutionary rates, we have tested various methods on the *mixture data* described in the Methods section. We present the performance of `Erik+2` (with $m = 1$ and $m = 2$), `MP`, `ML`(homGMc), and `ML` estimating a (heterogeneous across lineages) GTR model with discrete gamma rates with 2 categories, `ML`(GTR+2Γ) henceforth (Fig. 5.b). We included `MP` in this study because, as stated in Kolaczkowski and Thornton (2004), it performs better than `ML` estimating a single category model. This claim is confirmed by the results in our simulations with both versions of `ML`. It is worth pointing out that even `Erik+2` with $m = 1$ performs better than `ML`(homGMc) for internal branch length ≤ 0.25, and than `ML`(GTR+2Γ) for internal branch length ≤ 0.15. Also, notice that for length 10 000 and

larger, the accuracy of `Erik+2` with $m = 2$ is always greater than 33%, even if the internal branch length is small. This does not happen for `ML` or `MP`, which are not statistically consistent in this setting.

Performance on real data

We applied `ErikSVD`, `Erik+2` with $m = 1, 2, 3$, `ML` (GTR+2Γ+I), and `ML` (GTR+3Γ) to 4-taxon subalignments of the real data described above and investigated the proportion of output quartets that are compatible with T or T'. `Erik+2` and `ML` support the tree T and the model suggested by Jayaswal et al. (2014) ($m = 3$ for `Erik+2` and GTR+2Γ+I for `ML`), whereas `ErikSVD` gives more support to the alternative tree T' (Table 2).

We also represented the weights of each quartet output by `Erik+2` with $m = 3$ and `ML` (GTR+3Γ) (figure S6 in the Appendix 2). It is interesting to observe that `Erik+2` provides powerful discrimination between quartets and that the weights of the quartets that are incompatible with T are the lowest. Indeed, while `Erik+2` exhibits strong support for the correct topology and its weights vary from 0.3066 to 0.6700, `ML` weights are essentially equal and vary from .3314 to .3365. It is interesting to note that both methods tend to fail with the same quartets.

Execution time

We have compared the execution time of the different reconstruction methods used in our simulations with 100 alignments of length 1 000 bp. on a 3.2GHz processor. The results obtained show that `NJ` is the fastest method (1.324s), `ErikSVD` and `Erik+2` take 1.928s and 2.148s respectively, and `MP` takes 3.984s. Increasing the number m of categories in `Erik+2` does not increase running time (instead, it avoids some computations). Finally, `ML` is the slowest method by far because it has to infer the model parameters: on the same 100 alignments, `ML`(homGMc) and `ML`(homGTR) of PAML need about 10 seconds, and

ML(GTR) and ML(GTR+2Γ) of bppml (Bio++ package) need about 200 minutes (probably due to the large number of parameters)– we should point out that bppml allows many options and more experienced users may obtain better execution times.

Discussion and conclusions

The simulation studies show that Erik+2 is an accurate and robust topology reconstruction method on quartets, especially in situations where other methods systematically fail (model heterogeneity accross sites or lineages, or long-branch attraction). In such scenarios, Erik+2 outperforms the method of Eriksson, ErikSVD, and common methods like MP, NJ and ML based on models that cannot accommodate these assumptions. Erik+2 is based on the most general Markov model and hence accounts for heterogeneous data across lineages, even locally at each edge. When its assumptions are violated, for example in the presence of continuous gamma-distributed rates among sites, we have shown that it is highly accurate if there is enough data. As observed, Erik+2 can also deal with m-mixtures on the same tree topology (although for quartets the limit is $m = 3$). Even more, the simulations presented suggest that Erik+2 with $m = 2$ seems to be a good option to deal with large alignments when the mixture status of the data is unknown.

On the experiments we presented, the overall performance of ML is quite accurate if model assumptions are not violated, confirming the conclusions of Kolaczkowski and Thornton (2004); Kück et al. (2012). Also in line with these papers, we corroborate that long sequences do not improve ML performance on data that do not satisfy the hypothesis of the underlying model. Moreover, ML is by far the slowest among the methods tested here, while Erik+2 is not as fast as NJ, but still fast as its computing time is nearly twice NJ's. Another drawback of ML is that, quite often, it does not converge when it is computed on the incorrect topology, which makes the comparison of likelihoods impossible. Whereas the goal of Erik+2 is to reconstruct the topology, ML is designed to estimate the parameters

of the substitution matrices and it would probably be a good choice to use first `Erik+2` and then `ML` to estimate the parameters. In our simulation study, `NJ` (with paralinear distance) and `MP` have been the methods with least success, which is not so surprising if one takes into account that they are also less adaptable to general data. It is worth pointing out that in some cases (for example, short alignments), `ErikSVD` might be a better option than `Erik+2`, but in order to decide which method is best in each case, a detailed statistical study should be done.

We have only developed `Erik+2` for quartets with the aim of validating it as a successful method, and it is still a work in progress to further develop it for larger numbers of taxa. One possibility is to apply some quartet method based on the weights provided by `Erik+2`. Similar studies have been carried out by Rusinko and Hipp (2012) using weights obtained from phylogenetic invariants. In any case, the weighting method for the quartets plays a key role in this approach (Ranwez and Gascuel 2001). A different possibility would be to use `Erik+2` to evaluate the confidence of particular bipartitions of large sets of taxa, and in this case one can deal with a larger number m of categories (the maximum m allowed depends on the size of the subsets $\mathcal{A}$, $\mathcal{B}$ of taxa involved in the bipartition: $4^{min\{|\mathcal{A}|,|\mathcal{B}|\}-1} - 1$). However, some more work has to be done to adapt `Erik+2` to more taxa because the lack of data (short alignments relative to the size of the bipartition matrices) might affect the rank of the bipartition matrices. Extending `Erik+2` to amino acid data is also viable, and in this case the maximum number of allowed categories is equal to 19 for quartets (although we still have not addressed the possible natural biases that have to be considered in this case).

Acknowledgments

We are indebted to B. Misof and C. Mayer for encouraging us to pursue this project and for their very useful comments. We wish to thank them and their group for their warm

hospitality during our stay at the Alexander Koenig Zoological Museum. We also thank F. Anderson, B. Boussau, O. Gascuel, B. Holland, P. Jarvis, and L. Jermiin for useful comments and suggestions.

Finantial support

Both authors are partially supported by Spanish government MTM2012-38122-C03-01/FEDER and Generalitat de Catalunya 2009SGR1284, 2014SGR634.

Author's contributions

Both authors contributed equally to the development of this work.

*

References

Allman, E. S. and Rhodes, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.*, 13:1101–1113.

Allman, E. S. and Rhodes, J. A. (2007). Phylogenetic invariants. In Gascuel, O. and Steel, M. A., editors, *Reconstructing Evolution*. Oxford University Press.

Allman, E. S. and Rhodes, J. A. (2008). Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.*, 40(2):127–148.

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

Casanellas, M. and Fernández-Sánchez, J. (2007). Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees. *Mol. Biol. Evol.*, 24:288–293.

Casanellas, M. and Fernández-Sánchez, J. (2010). Relevant phylogenetic invariants of evolutionary models. *J. Math. Pure. Appl.*, 96:207–229.

Casanellas, M., Fernández-Sánchez, J., and Kedzierska, A. (2012). The space of phylogenetic mixtures for equivariant models. *Algorithms for Molecular Biology*, 7:33.

Casanellas, M. and Kedzierska, A. (2013). Generating Markov evolutionary matrices for a given branch length. *Linear Algebra and its Applications*, 438:2484-2499.

Cavender, J. A. and Felsenstein, J. (1987). Invariants of phylogenies in a simple case with discrete states. *J. Class.*, 4:57–71.

Demmel, J. W. (1997). *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Dutheil, J. and Boussau, B. (2008). Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evolutionary Biology*, 8(1):255.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Eriksson, N. (2005). Tree construction using singular value decomposition. In *Algebraic statistics for computational biology*, pages 347–358. Cambridge Univ. Press, New York.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Z*, 27:401–410.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.

Fitch, W. M. (1986). An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. *Progress in clinical and biological research*, 218:149–159.

Foster, P. G. (2004). Modeling compositional heterogeneity. *Syst. Biol.*, 53(3):485–495.

Galtier, N. and Gouy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of dna sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.*, 15:871–879.

Gu, X., and Li, WH. (1996). Bias-Corrected Paralinear and LogDet Distances and Tests of Molecular Clocks and Phylogenies Under Nonstationary Nucleotide Frequencies *Mol. Biol. Evol.*, 13(10):1375-1383.

Ho, S. Y. and Jermiin, L. S. (2004). Tracing the decay of the historical signal in biological sequence data. *Systematic Biology*, 53(4):623–637.

Holland, B. R., Jarvis, P. D., and Sumner, J. G. (2013). Low-parameter phylogenetic inference under the general Markov model. *Systematic Biology*, 63(1):78–92.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.*, 44:17–48.

Jayaswal, V., Jermiin, L. S., Poladian, L., and Robinson, J. (2011). Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. *Systematic Biology*, 60(1):74–86.

Jayaswal, V., Jermiin, L. S., and Robinson, J. (2005). Estimation of phylogeny using a general Markov model. *Evolutionary Bioinformatics Online*, 1:62–80.

Jayaswal, V., Wong, T. K., Robinson, J., Poladian, L., and Jermiin, L. S. (2014). Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic Biology*, 63(5):726–742.

Kedzierska, A. M. and Casanellas, M. (2012). Gennon-h: Generating multiple sequence alignments on nonhomogeneous phylogenetic trees. *BMC Bioinformatics*, 13(1):216.

Kolaczkowski, B. and Thornton, J. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431:980–984.

Kück, P., Mayer, C., Wagele, J.-W., and Misof, B. (2012). Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS One*, 7(10). DOI 10.1371/journal.pone.003

Lake, J. A. (1987). A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.*, 4:167–191.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proceedings of the National Academy of Sciences*, 91:1455–1459.

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, 21(7):1455–1458.

Rambaut, A. and Grassly, N. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13:235–238.

Ranwez, V. and Gascuel, O. (2001). Quartet-based phylogenetic inference: Improvements and limits. *Molecular Biology and Evolution*, 18(6):1103–1116.

Rhodes, J. A. and Sullivant, S. (2012). Identifiability of large phylogenetic mixture models. *Bulletin of Mathematical Biology*, 74(1):212–231.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425.

Rusinko, J. and Hipp, B (2012). Invariant based quartet puzzling *Algorithm Mol. Biol.*, 7(1):35

St. John, K., Warnow, T., Moret, B. M. E., and Vawter, L. (2003). Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining. *J. Algorithms*, 48(1):173–193.

Stefankovic, D. and Vigoda, E. (2007). Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Systematic biology*, 56(1):113–24.

Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13:964–960.

Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O., and Rogers, J. S. (2001). Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology*, 50(4):525–539.

Tateno, Y., Takezaki, N., and Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Molecular biology and evolution*, 11(2):261–77.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6):1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39:306–314.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13:555–556.

Yang, Z. and Roberts, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.*, 12:451–458.

Tables

Average success of different quartet methods on the tree space of Figure 1b.

simulations	base pairs	`ErikSVD`	`Erik+2`	`NJ`	`ML`
GM	1 000	**0.856** (0.21)	0.803 (0.17)	0.797 (0.18)	0.736 (0.17)
	10 000	0.958 (0.13)	**0.971** (0.04)	0.943 (0.09)	0.754 (0.17)
homGTR	500	0.732 (0.21)	0.748 (0.22)	0.729 (0.23)	**0.880** (0.11)
	1 000	0.796 (0.30)	0.843 (0.19)	0.805 (0.20)	**0.934** (0.06)
	10 000	0.940 (0.22)	**0.992** (0.04)	0.945 (0.10)	0.980 (0.02)

Table 1: Average success of `ErikSVD`, `Erik+2`, neighbor-joining (`NJ`) and maximum likelihood (`ML`) obtained in the results of Figure 3, 4, and Figure S2 in Appendix 2 (this is data simulated on the tree space of Figure 1b according to the general Markov model (GM) or the time-reversible model homogeneous across lineages and sites (homGTR) for different lengths). In parentheses we show the standard deviation of the set of percentages of success of each method in each tree space. In each row, the highest success is indicated in bold font. `ML`(homGMc) is applied when data are generated under the continuous GM model (that is, it estimates the most general homogeneous continuous-time model), while `ML`(homGTR) is applied when data are generated under the general time-reversible model homogeneous across lineages and sites.

Quartet compatibility of different methods with real data.

topology	`ErikSVD`	`Erik+2` ($m=1$)	`Erik+2` ($m=2$)	`Erik+2` ($m=3$)	`ML` (GTR+2Γ+I)	`ML` (GTR+3Γ)
T	91.43	**84.29**	**87.14**	**92.86**	**97.14**	**91.42**
T'	**94.26**	82.86	77.14	85.71	90.00	84.29

Table 2: Percentage of quartets output by `ErikSVD`, `Erik+2` (with different mixture assumptions), and `ML` (assuming a GTR model with either 2 discrete rate categories and invariable sites, or 3 discrete rate categories across sites) that are compatible with the yeast tree T of Rokas et al. (2003) and the alternative tree T' of Phillips et al. (2004). In each column, the highest success is indicated in bold font.

Figure captions

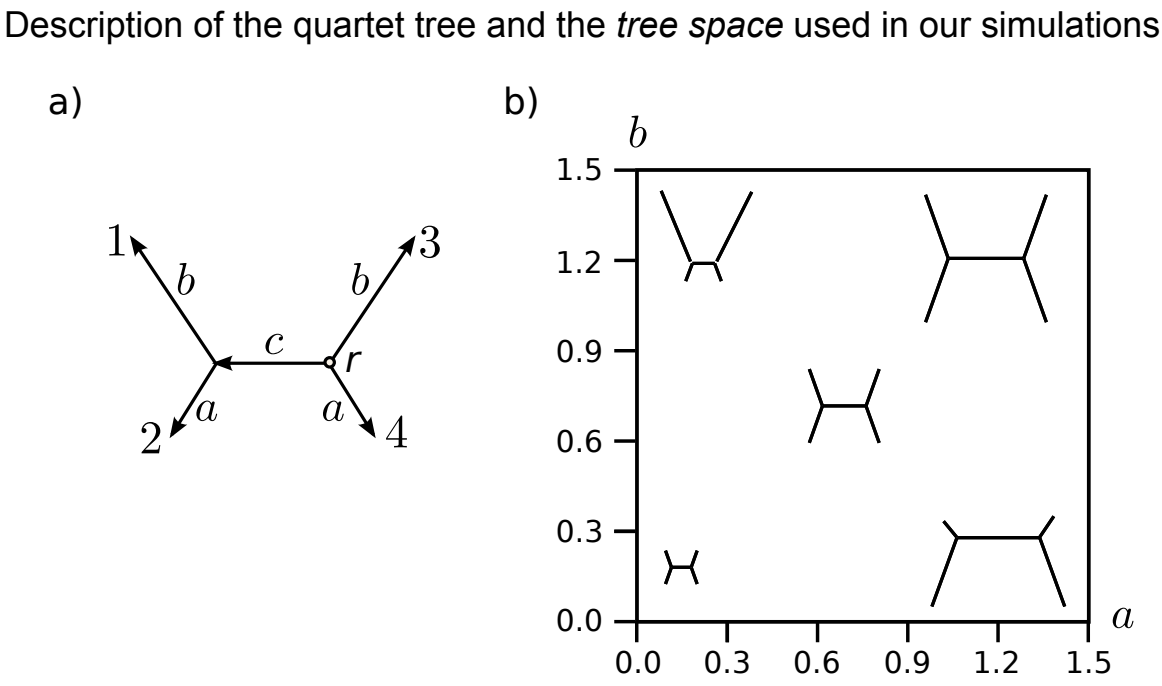

Figure 1: a) 4-leaf tree where the length of two opposite branches is represented by a; the other two peripheral branches have length b; and the length of the interior branch is denoted by c. The root r is located at the parent node of leaves 3 and 4. Branch lengths will be measured in the expected number of substitutions per site. b) Tree space obtained form the tree in a) when the branch length c is set equal to a and branch lengths a and b are varied from 0.01 to 1.5 in steps of 0.02 (see "Description of the data" in the Methods section).

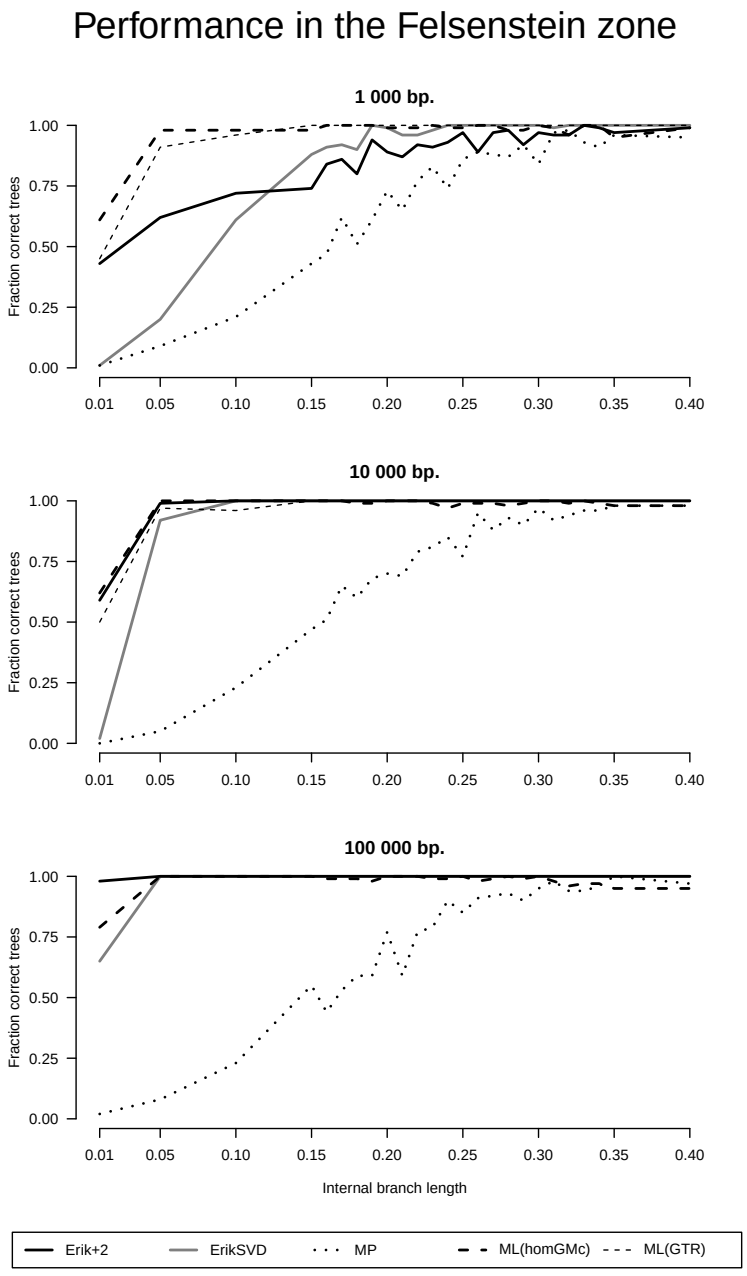

Figure 2: Percentage of correctly reconstructed topologies by `Erik+2`, `ErikSVD`, maximum likelihood `ML`, and maximum-parsimony `MP` on data generated under the general Markov model (GM) on the Felsenstein zone tree (see "Description of the data" in the Methods section). Two types of `ML` inference have been applied here: `ML`(homGMc) estimating the most general homogeneous continuous-time model, and `ML`(GTR) estimating a GTR model (due to the execution time of this last method, we could only test it for 1 000 bp. and 10 000 bp.).

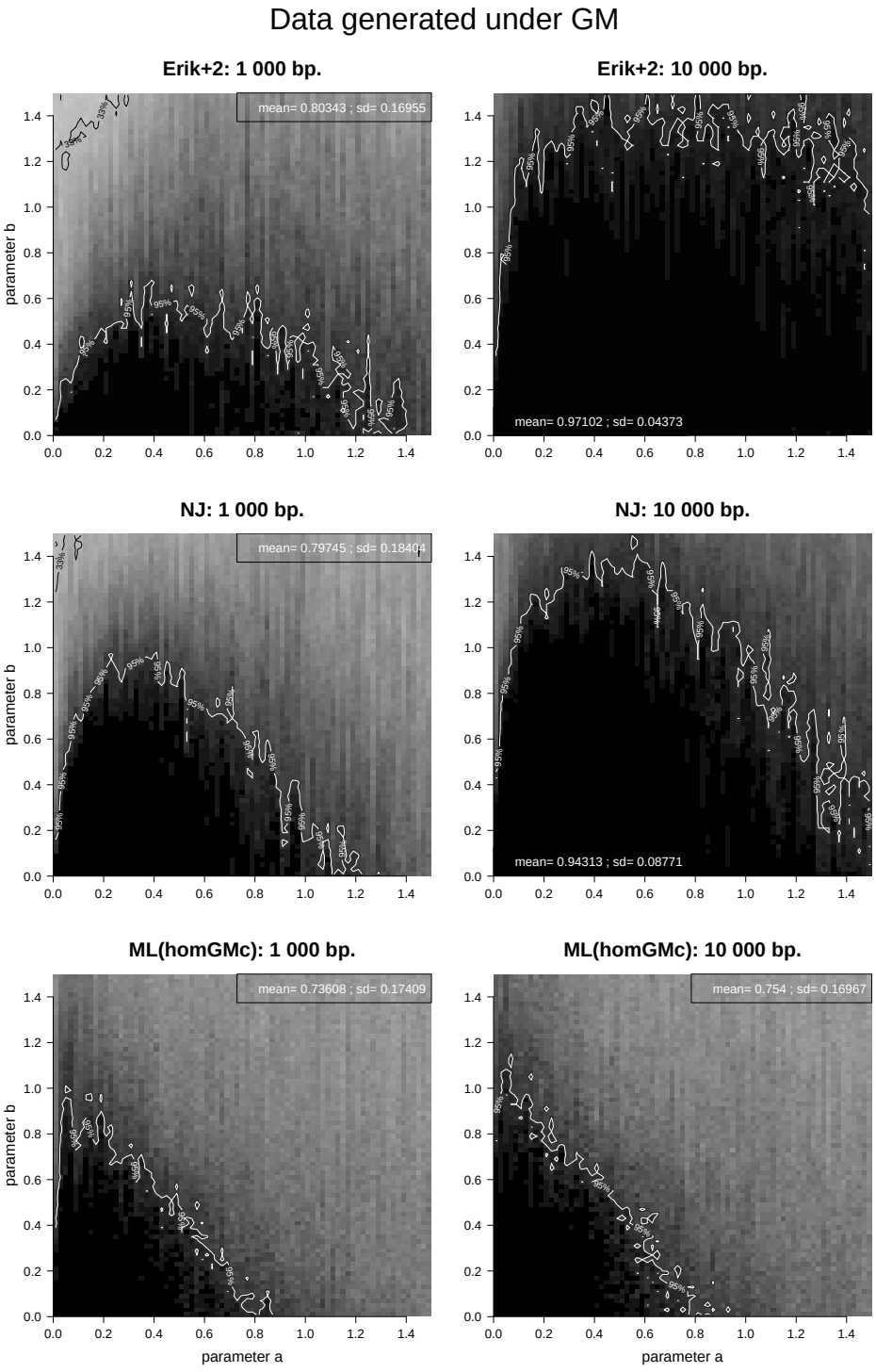

Figure 3: Performance in the tree space of Figure 1.b on data generated under the general Markov model *Left*: 1 000 bp; *Right*: 10 000 bp. Black is used to represent 100% of successful topology reconstruction, white to represent 0%, and different tones of gray the intermediate frequencies. The 95% contour line is drawn in white, whereas the 33% contour line is drawn in black. *Top*: `Erik+2`; *Middle*: neighbor-joining (paralinear distance); *Bottom*: `ML`(homGMc) estimating the most general homogeneous across lineages continuous-time model.

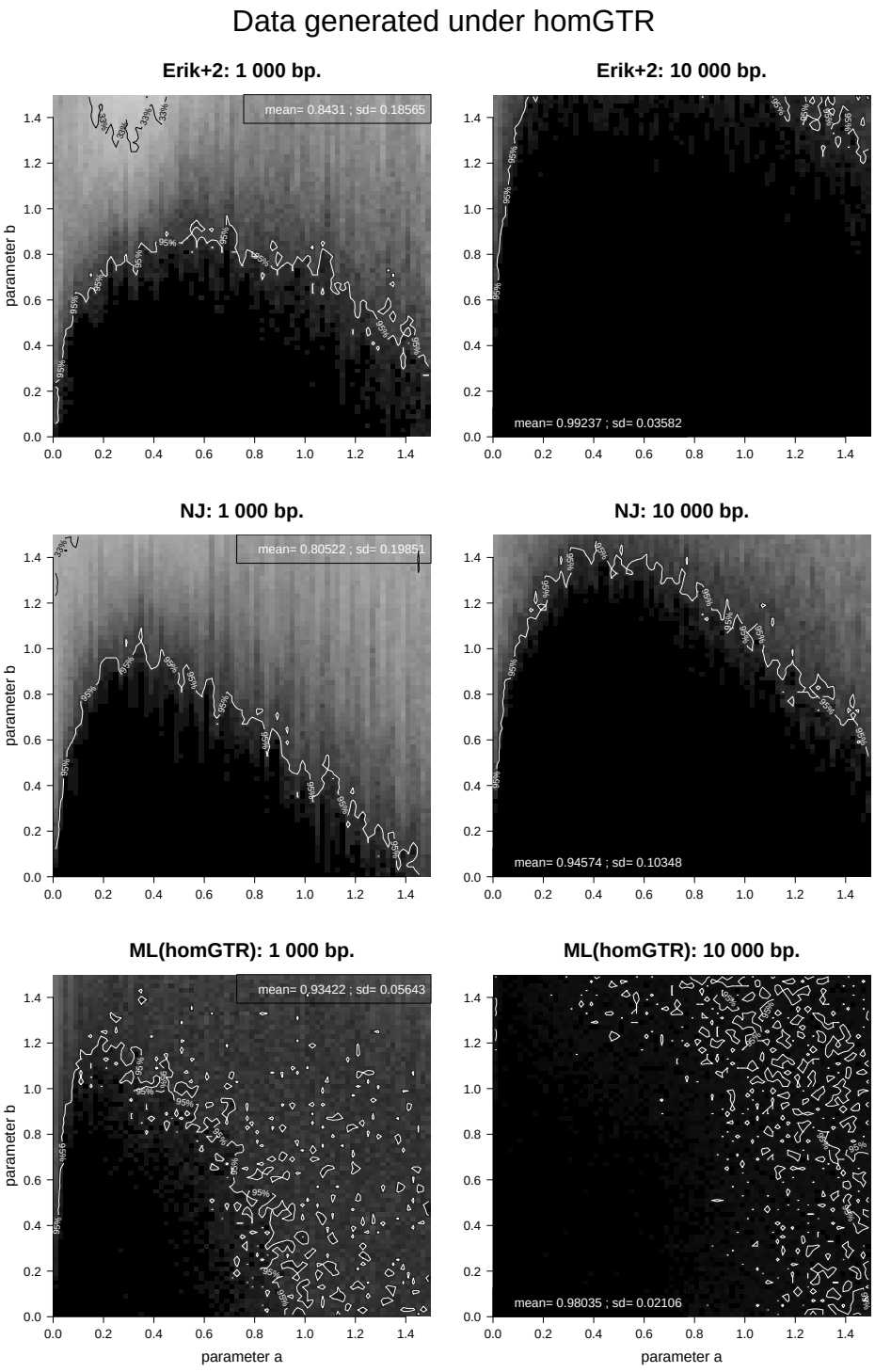

Figure 4: Performance in the tree space of Figure 1.b on data generated under the (homogeneous across lineages) GTR model *Left*: 1 000 bp; *Right*: 10 000 bp. Black is used to represent 100% of successful topology reconstruction, white to represent 0%, and different tones of gray the intermediate frequencies. The 95% contour line is drawn in white, whereas the 33% contour line is drawn in black. *Top*: `Erik+2`; *Middle*: neighbor-joining (paralinear distance); *Bottom*: `ML`(homGTR) estimating homogeneous GTR model.

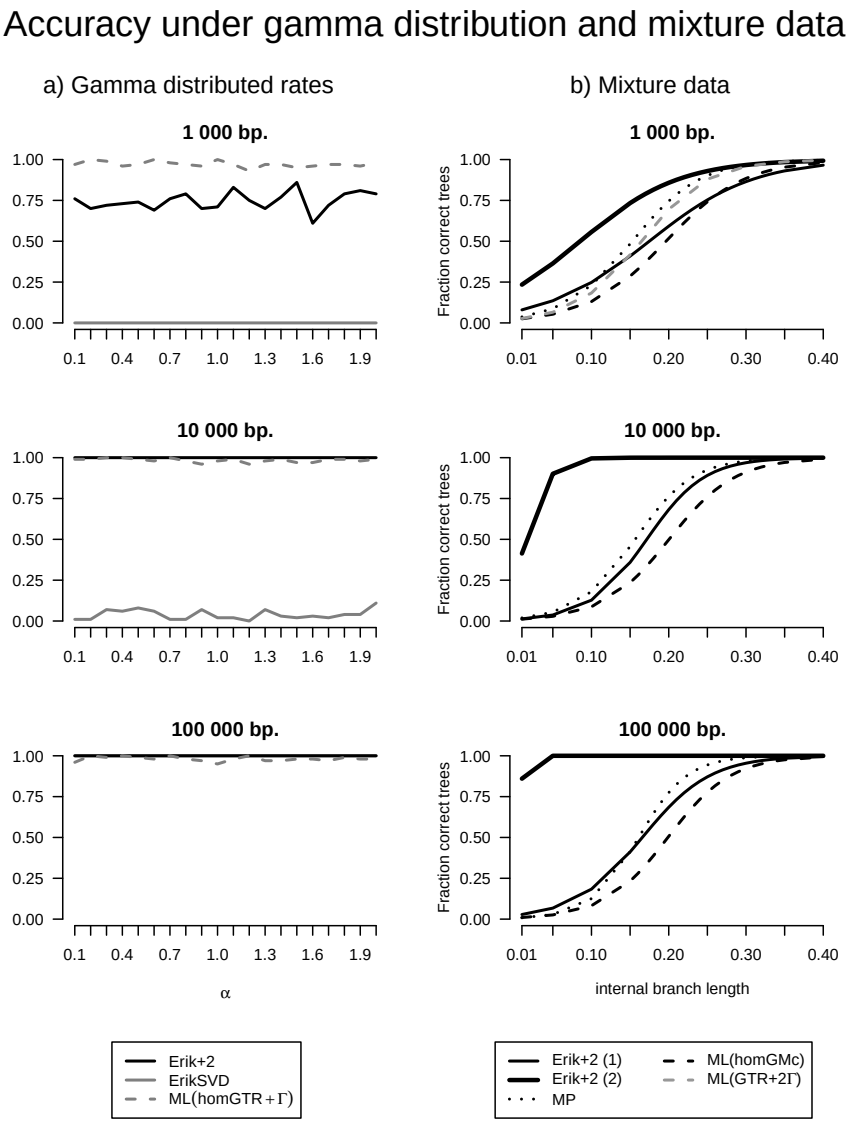

Figure 5: Percentage of correctly reconstructed topologies by different methods on alignments of lengths 1 000, 10 000 and 100 000 bp. shown from top to bottom. a) Data simulated under (homogeneous across lineages) GTR model with continuous gamma-rates and parameter α varying between 0.1 and 2 on the Felsenstein tree (see "Description of the data" in the Methods section). `Erik+2`, `ErikSVD`, and `ML` estimating the GTR model (homogeneous across lineages) with auto-discrete gamma-rates and denoted as `ML`(homGTR+Γ) (`MP` had 0% success on this data, so we do not show it). b) Data generated under the GM model with 2 categories according to the test designed in (Kolaczkowski and Thornton 2004), varying the internal branch length, and recovering with `Erik+2` with $m = 2$, `Erik+2` with $m = 1$, `MP`, `ML`(homGMc) estimating the most general homogeneous across lineages model, and `ML`(GTR+2Γ) estimating a time-reversible model with 2 discrete-gamma categories (due to the time of execution of this last method, we could only test it for 1 000 bp). The plot represents the logistic regression curve of the output of each method. In all cases, `ML` had to estimate all parameters.