SAQ: semi-algebraic quartet reconstruction method

Marta Casanellas, Jesús Fernández-Sánchez[1] and Marina Garrote-López, [1]

[1]*Dpt. Matemàtiques, Universitat Politècnica de Catalunya, Barcelona, Spain*

Universitat Politècnica de Catalunya, Av. Diagonal 647, 08028-Barcelona, Spain

Abstract.— We present the phylogenetic quartet reconstruction method SAQ (Semi-algebraic quartet reconstruction). SAQ is consistent with the most general Markov model of nucleotide substitution and, in particular, it allows for rate heterogeneity across lineages. Based on the algebraic and semi-algebraic description of distributions that arise from the general Markov model on a quartet, the method outputs normalized weights for the three trivalent quartets (which can be used as input of quartet-based methods). We show that SAQ is a highly competitive method that outperforms most of the well known reconstruction methods on data simulated under the general Markov model on 4-taxon trees. Moreover, it also achieves a high performance on data that violates the underlying assumptions.

[**Keywords**: pyhlogenetic reconstruction, general Markov model, quartet inference, algebraic phylogenetics]

E-mail addresses: marta.casanellas@upc.edu (M. Casanellas), jesus.fernandez.sanchez@upc.edu (J. Fernández-Sánchez), marina.garrote@upc.edu (M. Garrote-López)

Introduction

Phylogenetic reconstruction methods based on algebraic tools have been appointed as potential successful methods for gene reconstruction (Allman et al. 2017; Kubatko and Chifman 2019; Chifman and Kubatko 2015; Fernández-Sánchez and Casanellas 2016). These methods have the advantage of dealing with unrestricted underlying substitution models that allow for rate heterogeneity across lineages. However, they seem to require a large amount of data. For example, the method `Erik+2` (Fernández-Sánchez and Casanellas 2016) was proven to achieve a high performance under the general Markov model of nucleotide substitution (and allowed also for rate heterogeneity among sites), but needed about 10 000 sites with four taxa to clearly outperform neighbor-joining (NJ) or maximum likelihood (ML). A positive outcome of the algebraic methods for topology reconstruction is that they do not need to estimate the continuous parameters of the underlying evolutionary model. Although they are not ready to be used directly on large trees yet, they could be an excellent input for quartet-based reconstruction methods (Ranwez and Gascuel 2001) if a high performance for quartets could be achieved.

In Casanellas et al. (2020) we proved that algebraic tools may not be enough when dealing with confusing data (e.g. short alignments or quartets with a short internal branch) and that the stochastic nature of the data must be also considered. Theoretically, this could be done via the semi-algebraic description of the general Markov model presented in Allman et al. (2012). Nevertheless, in the same way that algebraic conditions do not directly use phylogenetic invariants (see Allman and Rhodes (2007) for an introduction to phylogenetic invariants), the semi-algebraic conditions cannot be used straightforward and separately from the algebraic ones. Providing a quartet reconstruction method that combines both and achieves a high performance is the goal of this article.

We present the method `SAQ`, which stands for Semi-Algebraic Quartet reconstruction method. It is a phylogenetic reconstruction method for DNA alignments on four

taxa which assumes a general Markov model of nucleotide substitution (GM henceforth) and which is based on the algebraic and semi-algebraic conditions that characterize data from this model. The underlying model is therefore the most general model of nucleotide substitution on independently and identically distributed sites (unrestricted distribution at the root, unrestricted transition matrices and rate heterogeneity across lineages). The method also outputs normalized quartet weights that can be used as input of quartet-based methods. Note that `SAQ` only aims to reconstruct the topology of the tree, not the substitution parameters.

We have tested `SAQ` on simulated data under different settings: on a "tree space" of quartets, on quartets with random branch lengths, and on alignments that are not identically distributed across sites ("mixture data" that violates the assumptions underlying `SAQ`). The data have been generated both under the GM model and a (homogeneous across lineages and) time-reversible continuous-time process (homGTR). Our results show that the method is highly successful, even with short alignments and with data that violates the assumptions of the method. The weights ouput by SAQ are shown to be unbiased and statistically consistent. We also provide a comparison of this method against existing methods such as ML, NJ and `Erik+2` and show that `SAQ` largely outperforms all of them for GM data and has a compatible performance for GTR data. Moreover, the results on mixtures of distributions on the same quartet show that the method is also able to deal with heterogeneity across lineages, as it surpasses methods that are specially designed for these data.

Methods

Description of the method

The method of phylogenetic reconstruction proposed in this paper is based on the theoretical result by Allman et al. (2012) that we briefly explain here. Given four taxa denoted by $\{1, 2, 3, 4\}$ we consider a DNA alignment of length N and collect the observed relative frequencies of site patterns as a vector in $\mathbb{R}^{256}$. The coordinates of a vector $p \in \mathbb{R}^{256}$ are labelled by $x_1x_2x_3x_4$, where each $x_i \in \{A, C, G, T\}$ stands for the observation at leaf i.

The set of the three fully-resolved (unrooted) quartet trees on the set of taxa is denoted as $\tau = \{12|34, 13|24, 14|23\}$. Given a bipartition $ij|kl$ on the set of taxa and a vector $p \in \mathbb{R}^{256}$, we denote by $F_{ij|kl}$ the *flattening* of p according to that bipartition: $F_{ij|kl}$ is the 16×16 matrix whose (x_ix_j, x_kx_l) entry is the coordinate of p that matches $x_ix_jx_kx_l$ in the convenient order (e.g. the (AC, GT) entry of $F_{12|34}$ is p_{ACGT} while the same entry in $F_{13|24}$ is p_{AGCT}).

Briefly speaking, the main result of Allman et al. (2012) states that a distribution $p \in \mathbb{R}^{256}$ comes from a general Markov (briefly GM) process on the quartet tree $T = 12|34$ if and only if the following three conditions are satisfied:

(a) the marginalization of p over each leaf comes from a Markov process on a tripod tree,

(b) the matrix $F_{12|34}$ has rank four (or less than four for special parameters),

(c) after applying sixteen "12|34 leaf-transformations" to p, the flattening matrices $F_{13|24}(\tilde{p})$ and $F_{14|23}(\tilde{p})$ associated to each transformed vector $\tilde{p}$ are symmetric and positive definite (or positive semidefinite for special parameters).

The first condition is independent of the tree topology, so it is not useful for recovering the tree topology. The second condition relies already on the tree topology and has been used in different phylogenetic reconstruction methods (Allman et al. 2017; Kubatko and Chifman 2019; Chifman and Kubatko 2015; Fernández-Sánchez and Casanellas 2016, see). However, this condition is satisfied by distributions arising not only from trees but also from certain networks (see Casanellas and Fernández-Sánchez (2020)) and, when restricted to trees, it does not reflect the stochasticity conditions of the transition matrices. The third condition reflects the stochastic nature of the transition matrix at the interior edge. For each quartet tree $T \in \tau$, there are sixteen "T leaf-transformations" that can be applied to any vector $p \in \mathbb{R}^{256}$. If p had arisen from a Markov process on a tree T' (not necessarily the same as T) with certain transition matrices, the resulting transformed vectors $\tilde{p}$ would have also arisen on T' but (in general) with different transition matrices. Although these transformations are fully explained in (Allman et al. 2012, see), we briefly illustrate them in the Appendix.

According to the results by Casanellas et al. (2020), it is important to combine conditions (b) and (c) in order to obtain successful reconstruction methods for data that might be misleading (that is, small samples or data coming from trees with a short interior edge).

In what follows, we explain how `SAQ` combines conditions (b) and (c). For a square matrix M, we denote by $psd(M)$ its closest positive semidefinite matrix (see Higham (1988)) and by $\delta_4(M)$ its distance to the set of rank ≤ 4 matrices Demmel (1997). The rank of $psd(M)$ is smaller than or equal to the rank of M (see Casanellas et al. (2018)). For each quartet tree T (in the set 12|34, 13|24, 14|23), given a distribution $p \in \mathbb{R}^{256}$, `SAQ` computes a score $s_T(p)$ as follows. If $T = 12|34$, we consider the sixteen 12|34 leaf-transformations $\tilde{p}_i$, $i = 1, \ldots, 16$ of p mentioned in (c) and for each of these vectors we

compute

$$s_T^i := \frac{\min\left\{\delta_4(psd(F_{13|24}(\tilde{p}_i))), \delta_4(psd(F_{14|23}(\tilde{p}_i)))\right\}}{\delta_4\left(psd(F_{12|34}(\tilde{p}_i))\right)}.$$

Then, define $s_T(p)$ as the average of these sixteen quantities. If T is any of the other two trees, $s_T(p)$ is computed analogously by permuting the roles of the leaves accordingly. Finally `SAQ` outputs the normalized three scores, that is, if $s := s_{12|34}(p) + s_{13|24}(p) + s_{14|23}(p)$, then

$$\texttt{SAQ}(p) := \frac{1}{s}\left(s_{12|34}(p), s_{13|24}(p), s_{14|23}(p)\right).$$

If p arises as a distribution on the tree $12|34$ with stochastic parameters, then $F_{12|34}(\tilde{p}_i)$ has rank ≤ 4 (by (b)) and $psd(F_{12|34}(\tilde{p}_i))$ has also rank ≤ 4 , so $\delta_4(psd(F_{12|34}(\tilde{p}_i))) = 0$. Moreover, if p comes from generic parameters (namely, invertible transition matrices with positive entries and postive distribution at the root node), then $F_{13|24}(\tilde{p}_i)$ (resp. $F_{14|23}(\tilde{p}_i)$) is a symmetric positive definite matrix by (c) and has rank 16 (Allman et al. 2012, proof of Proposition 5.6), so actually $\delta_4(psd(F_{13|24}(\tilde{p}_i))) = \delta_4(F_{13|24}(\tilde{p}_i)) > 0$ (analogously for $F_{14|23}(\tilde{p}_i)$). Therefore, $s_{12|34}(q) \to \infty$ when q approaches a distribution p that was generated on the tree $12|34$ (with generic stochastic parameters).

On the other hand, $s_{13|24}(q)$ and $s_{14|23}(q)$ tend to zero when q approaches a distribution p generated on $12|34$ with generic stochastic parameters. Indeed, in order to compute $s_{13|24}(p)$ we consider the sixteen $13|24$ leaf-transformations of p, which will be denoted as $\hat{p}_i$; then $F_{12|34}(\hat{p}_i)$ has still rank ≤ 4 (as mentioned above, $\hat{p}_i$ still arises from $12|34$) and its closest positive definite matrix must have rank ≤ 4 (Casanellas et al. 2018). Thus, $\min\{\delta_4(psd(F_{12|34}(\hat{p}_i))), \delta_4(psd(F_{14|23}(\hat{p}_i)))\} = 0$. Moreover, if p arises from generic parameters, $F_{13|24}(\hat{p}_i)$ has rank 16, is not symmetric anymore and its closest positive definite matrix has generically rank strictly larger than four (see a justification in the Appendix). Hence, in this case $\delta_4(psd(F_{13|24}(\hat{p}_i))) > 0$ and $s_{13|24}(p)$ is zero if p comes from generic parameters on the tree $12|34$. A similar argument applies to $14|23$.By normalizing the scores we get that, if $q \in \mathbb{R}^{256}$ is a distribution that tends to a distribution p generated

on the tree 12|34 with generic stochastic parameters, then

$$\lim_{q \to p} \texttt{SAQ}(q) = \texttt{SAQ}(p) = (1, 0, 0).$$

According to the `SAQ` method, the correct topology for a distribution p is the topology T for which $s_T(p)$ attains the maximum value. Hereby, `SAQ` seeks to minimize the average distance of the flattening F_T of the transformations to rank four matrices and to maximize the average distance of the other two flattenings to rank four, under the assumption that these should be positive definite. SAQ can be understood as a quartet inference measure in the sense of Sumner et al. (2017).

In practice, the implementation of an algorithm that computes these scores requires dealing with some technical difficulties. For example, the so-called leaf transformations might require computing the inverse of ill-conditioned matrices derived from the marginalization over two taxa. Moreover, when p is an empirical distribution, then its leaf transformations are not distributions any more. The implementation we provide in

`https://github.com/marinagarrote/SAQ-method`.

excludes leaf transformations that are far from being distributions (this is the parameter "filter" that can be modified by the user and that could be adapted to the alignment length).

Description of the simulated data

In order to test the new method `SAQ` and compare its performance to other reconstruction methods, we have used the data of Fernández-Sánchez and Casanellas (2016). These data had been generated under two different models, the general Markov model (GM) and a time-homogenous GTR model (homGTR), under a wide range of systems of branch lengths on quartet trees. We briefly describe these sets of simulated data and refer the reader to Fernández-Sánchez and Casanellas (2016) for more details. Branch

lengths here are measured as the expected number of elapsed substitutions per site along the evolutionary process associated to the branch.

Models of nucleotide substitution.— In reference to the general Markov model (GM), the Markov process on the tree starts with a random distribution and evolves according to random Markov matrices at the edges that correspond to the given branch length. These data had been generated using the software `GenNon-h` (Kedzierska and Casanellas 2012).

By a time-homogeneous GTR model (homGTR) we mean a continuous-time GTR model that shares the same instantaneous mutation rate matrix Q at all branches of the tree. These data had been generated using `Seq-gen` (Rambaut and Grassly 1997).

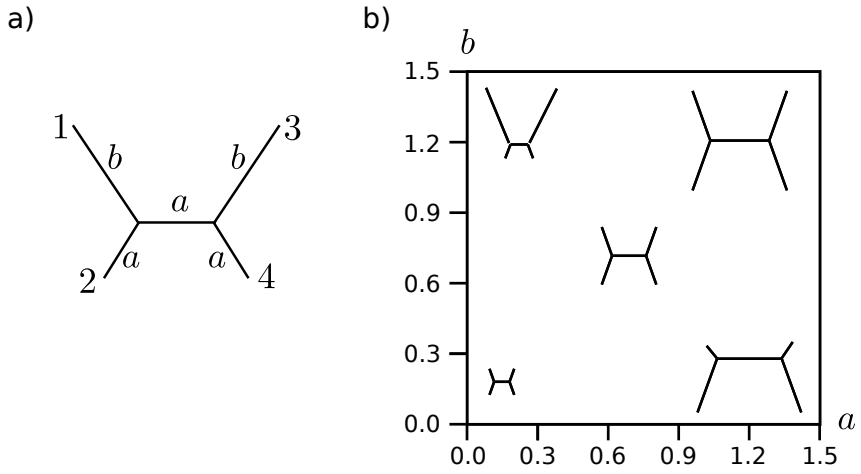

Figure 1: a) 4-leaf tree where the length of two opposite branches and the interior branch are represented by a; the other two peripheral branches have length b; is denoted by c. Branch lengths will be measured in the expected number of substitutions per site. b) Tree space obtained form the tree in a) when the branch lengths a and b are varied from 0.01 to 1.5 in steps of 0.02.

Tree space.— We use the parameter space suggested in Huelsenbeck (1995) to test different methods of phylogenetic reconstruction. More precisely, we evaluate the method on a *tree space* (see Figure 1.b) where the quartets are as in Figure 1.a, and the branch lengths a and b vary between 0 and 1.5 in steps of 0.02. Note that on the top left part of this

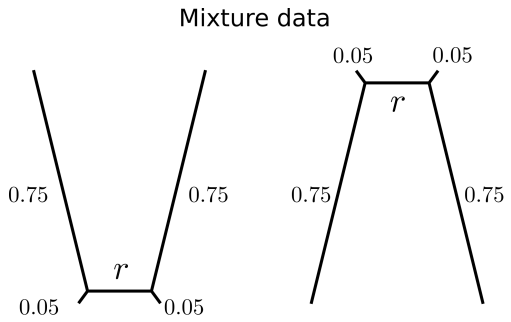

Figure 2: Mixture data taken from Kolaczkowski and Thornton (2004): two categories of the same size are considered, both evolving under the GM model on the two trees depicted above with the branch lenghts indicated. The internal branch length takes the same value in both categories and varies from 0.01 to 0.4 in steps of 0.05.

tree space we find the "Felsenstein zone", which contains trees sbuject to the long branch attraction phenomenon. For each pair (a, b), we have one hundred alignments generated under this setting for each of the two models of nucleotide substitution considered, GM and hom-GTR.

Random trees.— A total of 10 000 alignments are considered, obtained from 4-taxa trees with random branch lengths uniformly distributed either in the interval (0,1) or in (0,3), and generated according to one of the two substitution models, GM or homGTR.

Mixture data.— In order to test the performance of `SAQ` when the assumptions underlying the method are violated, we have taken the approach by Kolaczkowski and Thornton (2004) and considered mixtures of distributions as follows. We take two categories of the same sample 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 $b = 0.05$ (see Figure 2). The internal branch length takes the same value in both categories and varies from 0.01 to 0.4 in steps of 0.05 (see figure 2).

Results

Tree space

The performance of `SAQ` on the tree space data for alignments of 500 and 1 000 bp. simulated under GM or homGTR data is presented in figure 3. The success of `SAQ` in recovering the correct quartet is represented by different tones of gray, where black corresponds to a 100% success and white corresponds to a 0% success. Gray tones correspond to regions of intermediate probability, and the 95 % and 33 % isoclines are represented with a white line.

We observe that these figures exhibit a consistent performance (according to the results by Huelsenbeck (1995) and Fernández-Sánchez and Casanellas (2016)), with a decreasing performance at the Felsenstein zone but a high performance at the other zones, and with an increase of success for larger samples. In summary, the average success of `SAQ` applied to alignments of length 10 000 bp. is 96.8 % when applied to GM data, and 94.5 % when applied to homGTR data.

In Table 1 we summarize the overall performance of `SAQ` on the tree space in comparison with the methods studied (Fernández-Sánchez and Casanellas 2016). In particular, we compare it with a maximum likelihood approach: ML(homGMc) estimates the most general homogeneous across lineages continuous-time model and ML(homGTR) estimates a homogeneous across lineages GTR model (both methods estimate the rate matrix and the distribution at the root). We also compare it to the neighbor-joining method (NJ) under the paralinear distance, as was used in the quoted paper. We also provide the comparison to the algebraic method `Erik+2`, which is based on a GM model and makes only use of condition (b) explained in the Methods section. The plots of the performance of these methods on the tree space described above can be found in Fernández-Sánchez and Casanellas (2016); here we summarize the results in Table 1.

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

simulations	base pairs	SAQ	Erik+2	NJ	ML
GM	500	84.6	72.4	72.5	72.1
	1 000	88.8	80.3	79.7	73.6
homGTR	500	78.4	74.8	72.9	88.0
	1 000	83.5	84.3	80.5	93.4

Table 1: Average success of SAQ corresponding to the tree space of Figure 1 b, compared with the results of the performance of Erik+2, neighbor-joining (NJ) and maximum likelihood (ML) taken from (Fernández-Sánchez and Casanellas 2016). ML(homGMc) estimates a homogeneous continuous GM model (that is, it estimates an unrestricted rate matrix for the whole tree and a distribution at the root) and is applied when data is generated under a GM model, while ML(homGTR) is applied when data are generated under homGTR.

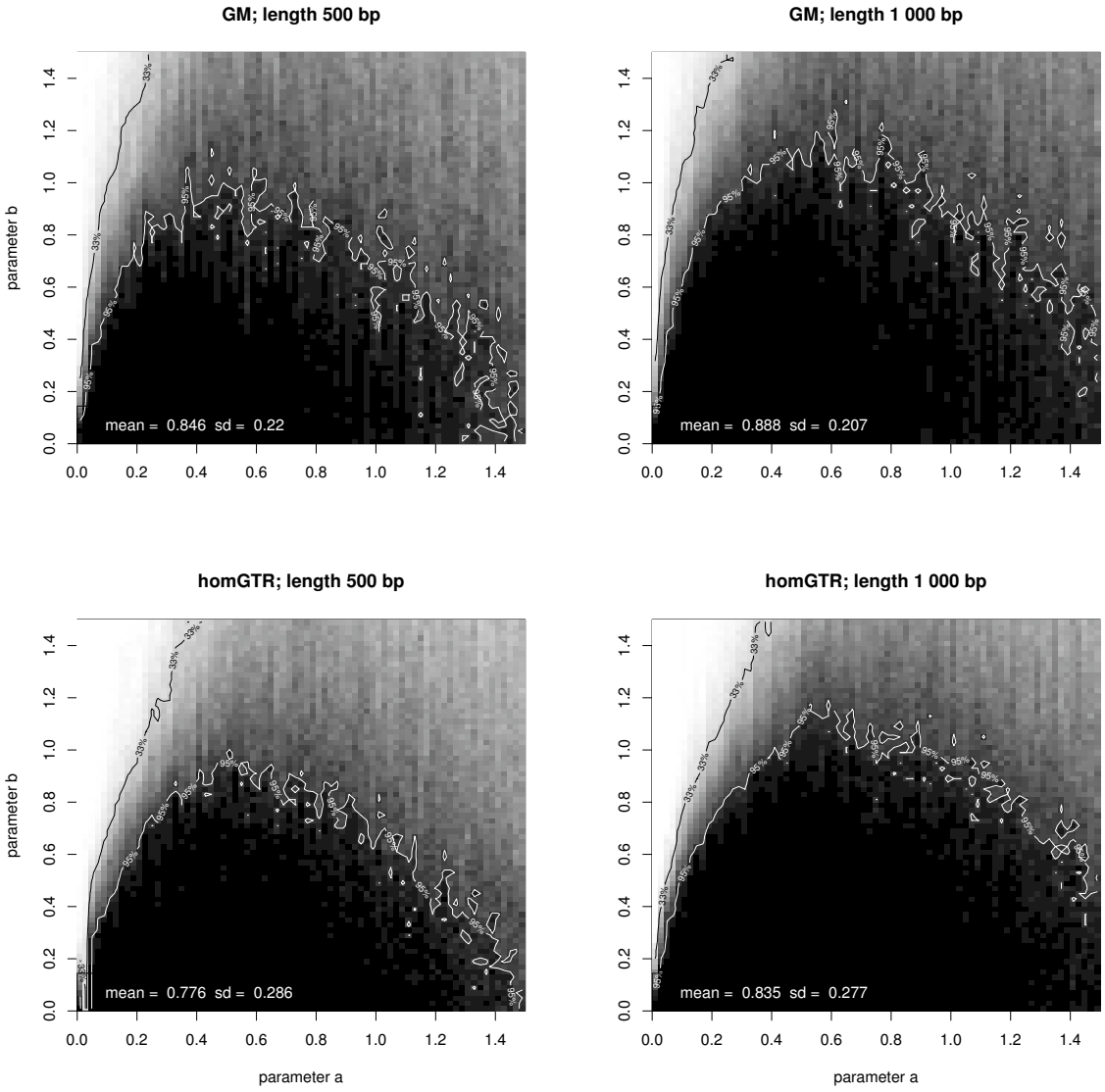

Figure 3: Performance of `SAQ` in the tree space of Figure 1.b on alignments of length 500 bp (left) and 1 000 bp (right). Black is used to represent 100% of successful quartet reconstruction, white to represent 0%, and different tones of gray the intermediate frequencies. The 95% contour line is drawn in white, whereas the 33% contour line is drawn in black. Top: data generated under the GM model; Bottom: data generated under a homogenous GTR model.

Random branch lengths

As the weights of `SAQ` are normalized so that the three quartet values sum to one, it is suitable for input of quartet-based methods and also allows plotting the scores in a ternary plot (also called a simplex plot), see (Strimmer and von Haeseler 1997). To visualize how the output of `SAQ` is distributed, we show ternary plots corresponding to the alignments generated on the tree 12|34 under the GM and homGTR models, with lengths 1 000 bp. and 10 000 bp. and random branch lengths uniformly distributed in (0,1) (Figure 4). The ternary plots of the performance of `SAQ` when applied to the same setting with branch length uniformly distributed in (0,3) is shown in Figure 7 of the Appendix.

We observe that `SAQ` weights are equally distributed for the two wrong topologies and the vast majority of points lie close to the left corner (which represents the correct quartet) for both GM and homGTR data. We also observe a strong difference in the performance of `SAQ` when applied to branch lengths in (0,1) or in (0,3), being notable higher in the former. The average success of `SAQ` for these random branches systems is shown in Table 2.

Average success of `SAQ` applied to data generated on 12|34 with random branch lengths

	GM		homGTR	
	1 000 bp.	10 000 bp.	1 000 bp.	10 000 bp.
(0,1)	95.74	98.72	93.72	98.06
(0,3)	69.74	82.69	67.87	84.51

Table 2: Average success of `SAQ` on alignments of lengths 1000 and 10000 bp. generated on the tree 12|34 under the GM and homGTR models with random branch lenghts uniformely distributed in (0,1) (first row) and (0,3) (second row).

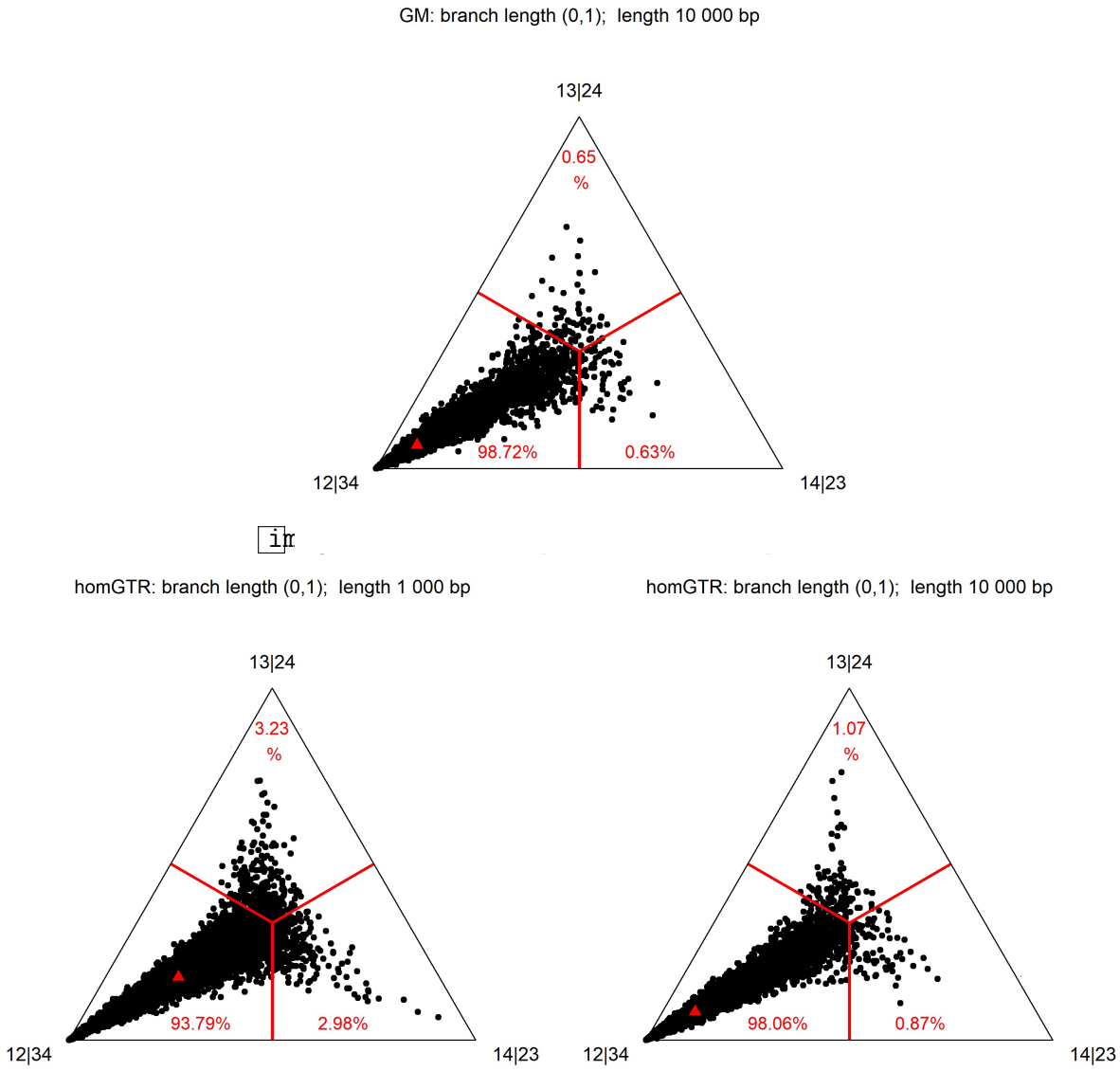

Figure 4: Ternary plots corresponding to the weights of `SAQ` applied to 10000 alignments generated under the 12|34 tree with random branch lengths uniformly distributed between 0 and 1. On each triangle the bottom-left vertex represents the underlying tree 12|34, the bottom-right vertex is the tree 13|24 and the top vertex is 14|23. Top: correspond to data generated under GMM; bottom: data generated under GTR. Left : 1000 bp; Right: 10000 bp.

Mixture data

The performance of the method `SAQ` under the mixed data described in the Methods section (see also figure 2) is shown in figure 5. The ternary diagrams show a high accuracy in determining the correct quartet, even when the length of the alignments is 1 000 bp. In the same figure, we present the performance of `SAQ` in terms of the branch length of the interior edge, following the study suggested by Kolaczkowski and Thornton (2004), for lengths 1 000, 10 000 and 100 000 bp. This plot is to be compared with the analogous plot in Fernández-Sánchez and Casanellas (2016) and we summarize the comparision to other methods in Table 3. As it is apparent from the figures values? results? in the table, `SAQ` outperforms all the other methods (even `Erik+2` (2)) despite mixture data violates the assumptions of this method.

Performance of different methods applied to mixture data

internal branch length	0.01	0.05	0.1	0.2	0.3
`SAQ`	37	83	96	100	100
`Erik+2 (2)`	12	35	60	86	96
`MP`	0	2	19	76	99
`ML`(GTR+2 Γ)	0	4	14	77	95

Table 3: Percentage of correctly reconstructed topologies by different methods on alignments of length 1 000 bp. for data generated under the GM model with 2 categories according to the test designed described in the section data and varying the internal branch length, and recovering with `SAQ`, `Erik+2` with 2 partitions, Maximum Parsimony and `ML`(GTR+2 Γ) estimating time-reversible model with 2 discrete-gamma categories. For all internal branch lenghts, ML had to estimate all parameters.

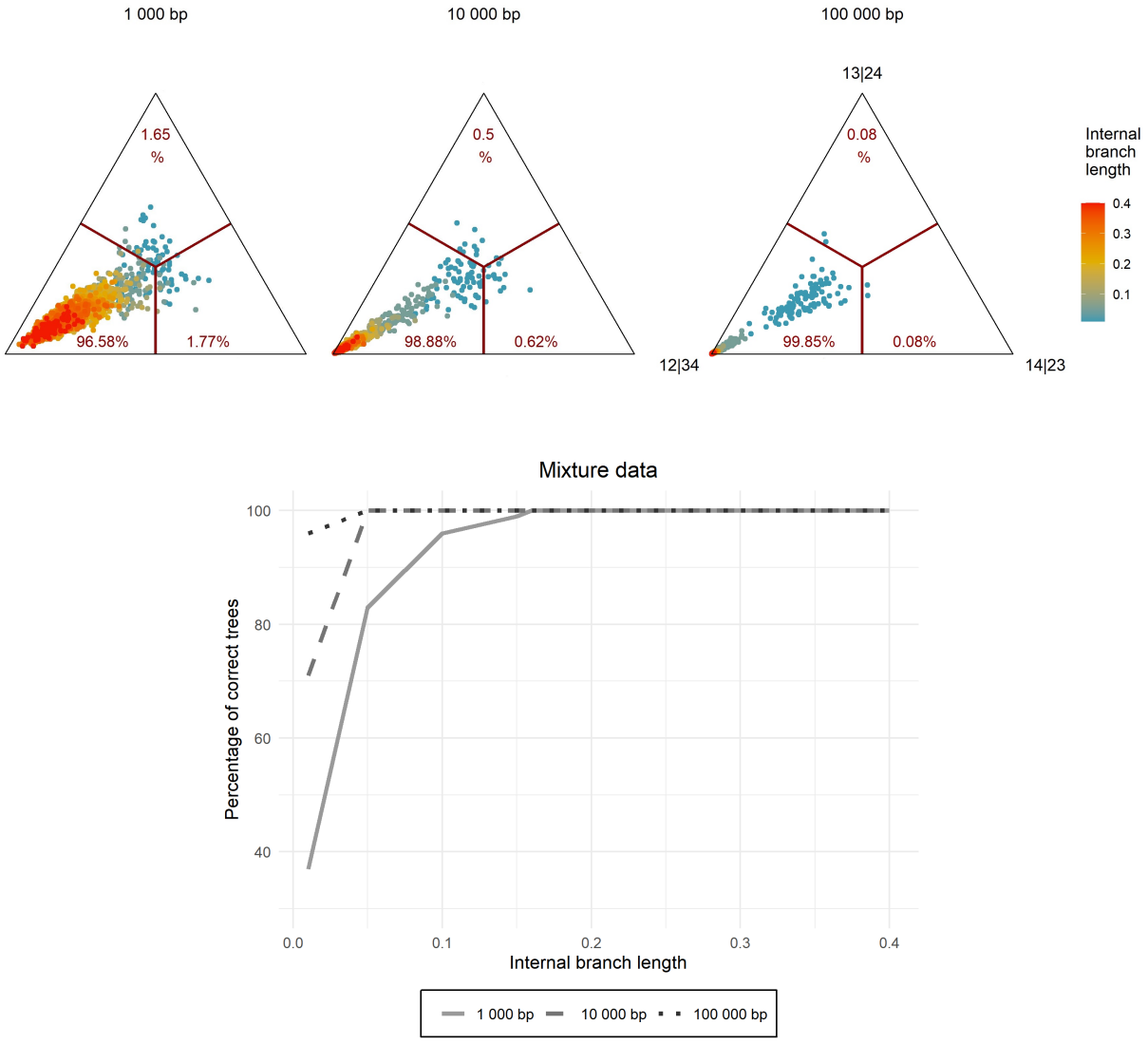

Figure 5: The three plots on the top show the ternary diagrams corresponding to the weights of `SAQ` applied to heterogeneous alignments of lengths 1 000, 10 000 and 100 000 bp. generated by the trees of figure 2. On each triangle the bottom-left vertex represents the underlying true topology 12|34, the bottom-right vertex is topology 13|24 and the top vertex is 14|23. Each triangle is divided into three regions according to which tree is selected by the method. The figures in red represent the percentage of alignments of the corresponding region according to `SAQ`. The gradient in color of the dots describe the variation of the branch length r of the interior edge, from 0.01 to 0.4 in steps of 0.05, as indicated by the bar on the right. For the same data, the plot on the bottom represents the percentage of correctly reconstructed trees by `SAQ` as a function of the internal branch lenght. Data generated under the GM model with 2 categories, varying the internal branch length, and recovering with `SAQ`.

Execution time

The computations were performed on a machine with 6 Dual Core Intel(R) Xeon(R) E5-2430 Processor (2.20 GHz) equipped with 25 GB RAM running Debian GNU/Linux 8. We have used the g++ (Debian 4.9.2-10+deb8u2) 4.9.2 compiler and the C++ library for linear algebra & scientific computing *Armadillo* version 3.2.3 (Creamfields).

The time needed to apply `SAQ` to 100 alignments of length 10 000 bp is 9 seconds.

Discussion and conclusions

In this paper, we have presented a new quartet reconstruction method that we call `SAQ`. It is a robust and accurate method, which is essentially based on the results obtained by Allman et al. (2012) and makes profit of the stochastic information available in the data to infer the topology of the phylogenetic tree. As far as we are aware, this is the first method that combines both the algebraic and the semialgebraic nature of the underlying substitution models.

The study and the simulations carried out in this paper show that `SAQ` is a robust phylogenetic reconstrucion method, specially when dealing with data generated under the GM model. We have even proven that the performance of the method is high for data that violate the hypothesis of the model (mixture data). In connection with the performance on mixed data, we would like to point out that although `SAQ` is not specifically designed to deal with mixtures of distributions in general, it can be easily adapted to deal with invariable sites. One only needs to switch the relative frequency vector p by a corrected vector that takes into account the estimated amount of invariable sites (e.g. as in Jayaswal et al. (2007) or Steel et al. (2000)). An implementation of this feature in `SAQ` is on the way.

The results described in the preceeding section show that the method performs better when applied to data generated under the GM model than under a homogeneous

GTR model (for which the results are satisfactory and similar to the method `Erik+2`). One possible explanation for this phenomena is that on GTR data, when the transition matrix at the interior edge is close to the identity (so that the method is pushed to the limit), the flattening matrices $F_{12|34}$ needed to compute the weights of `SAQ` are close to be symmetric and positive definite (at least when the stationary distribution is uniform). In this situation, the semialgebraic information of the data become almost irrelevant and the method practically relies on the distance to 4-rank matrices. As a consequence, the results are similar to the results of other methods based on this distance, as `Erik+2`.

`SAQ` is developed for quartets and assigns a normalized weight between 0 and 1 to each possible tree: this weight is larger to the tree with the higher confidence. We have checked that these normalized weigths are unbiased (in the sense that there is no trend towards any of the incorrect trees) and are statistically consistent. Although `SAQ` is only developed for quartets, its weights can be used as input of quartet-based methods such as "weight optimization" by Ranwez and Gascuel (2001), "quartet-puzzling" by Strimmer and von Haeseler (1996) or the method by Willson (1999). We plan to test the weights of `SAQ` as input of quartet-based methods in a forthcoming work.

Finally, let us note that combining algebraic and semi-algebraic conditions in a single score has been a tedious task. There are many different ways of combining all the restrictions satisfied by the theoretical distributions. For example, in the definition of the score s^i_T, we could have considered not projecting into symmetric positive matrices for the flattening $F_{12|34}$ as this seems as adding unnecessary complexity to the algorithm. However, we found that by considering the projection, all distances to 4-rank matrices are computed on the same space of positive definite matrices and improves the results of the algorithm. Analogously, we have tested many other options that are not mentioned in this final report. In this sense, `SAQ` is the result of a number of decisions based on exhaustive simulation studies, with the scope of obtaining a method that takes into account all the information at hand, but at the same time, keeping the method computationally feasible

and as simple as possible. We are aware that a detailed statistical analysis would be convenient in order to justify the good performance observed in the simulations.

Finantial support

The authors were partially supported by Spanish government Secretaría de Estado de Investigación, Desarrollo e Innovación [MTM2015-69135-P (MINECO/FEDER)] and [PID2019-103849GB-I00 (MINECO)]; Generalitat de Catalunya [2014 SGR-634]. M. Garrote-López was also funded by Spanish government, Ministerio de Economía y Competitividad research project Maria de Maeztu [MDM-2014-0445].

Author's contributions

All authors contributed equally.

*

References

Allman, E. S., Kubatko, L. S., and Rhodes, J. A. (2017). Split Scores: A Tool to Quantify Phylogenetic Signal in Genome-Scale Data. *Systematic Biology*, 66(4):620–636.

Allman, E. S. and Rhodes, J. A. (2004). *Mathematical models in biology, an introduction.* Cambridge University Press. ISBN 0-521-52586-1).

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., Rhodes, J. A., and Taylor, A. (2012). A semialgebraic description of the general Markov model on phylogenetic trees. *SIAM Journal on Discrete Mathematics*, 28.

Carrell, J. B. (2017). *Groups, Matrices, and Vector Spaces*. Springer New York.

Casanellas, M. and Fernández-Sánchez, J. (2020). Rank conditions on phylogenetic networks. In *Research Perspectives CRM Barcelona. Spring 2019*, volume 10 of *Trends in Mathematics*, page to appear. Springer-Birkhäuser.

Casanellas, M., Fernández-Sánchez, J., and Garrote-López, M. (2018). The inertia of the symmetric approximation for low rank matrices. *Linear and Multilinear algebra*, 66(11):2349–2353.

Casanellas, M., Fernández-Sánchez, J., and Garrote-López, M. (2020). Distance to the stochastic part of phylogenetic varieties. *Journal of Symbolic Computation*.

Chifman, J. and Kubatko, L. (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, 374:35–47.

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

Fernández-Sánchez, J. and Casanellas, M. (2016). Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. *Systematic Biology*, 65(2):280–291.

Ghys, Ã. and Ranicki, A. (2016). Signatures in algebra, topology and dynamics. *Ensaios MatemÃ¡ticos*, 30:1–173.

Higham, N. J. (1988). Computing a Nearest Symmetric Positive Semidefinite Matrix. *Linear Algebra and its Applications*, 103:103–118.

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

Jayaswal, V., Robinson, J., and Jermiin, L. (2007). Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution. *Syst. Biol.*, 56:155–162.

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

Kubatko, L. S. and Chifman, J. (2019). An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evolutionary Biology*, 19(112).

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.

Steel, M., Huson, D., and Lockhart, P. (2000). Invariable sites models and their use in phylogeny reconstruction. *Syst. Biol.*, 49(2):225–232.

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

Strimmer, K. and von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci.*, 94:6815–1819.

Sumner, J. G., Taylor, A. E., Holland, B. R., and Jarvis, P. D. (2017). Developing a statistically powerful measure for quartet tree inference using phylogenetic identities and markov invariants. *Journal of mathematical biology*, 75 6-7:1619–1654.

Willson, S. (1999). Building Phylogenetic Trees from Quartets by Using Local Inconsistency Measures. *Molecular Biology and Evolution*, 16(5):685–685.

Appendix

Insight on the theoretical basis of the method

T leaf-transformations.— Let p be a distribution arising from a Markov process on the tree $T = 12|34$ with transition matrices M_1, M_2, M_3, M_4 at the external edges and M_5 at the internal edge. The effect of the T leaf-transformations on p is that of replacing some of the external matrices so that both leaves in the same side of the tree share the same matrix. This is achieved in 4 possible patterns, namely $M_1|M_3$, $M_1|M_4$, $M_2|M_3$, or $M_2|M_4$ (see Figure 6, middle row). There are four possible ways of computing each of these transformed distributions, and these four ways are not equivalent when applied to a distribution q that has not arisen as a Markov process on T. Therefore, in general, we have 16 T leaf-transformations that can be applied to a distribution $q \in \mathbb{R}^{256}$.

Analogously, if $T' \neq T$, we also have 16 T' leaf-transformations that act on distributions in $\mathbb{R}^{256}$. Remarkably, if p is a distribution that arises from a Markov process on T, these T' leaf-transformations applied to p produce 16 vectors $\hat{p}_i$ that have also arisen on T but with different matrices at the exterior edges (see figure 6, bottom row). Moreover, while the former T leaf-transformations applied to p produce *distributions*, the latter may produce non-stochastic vectors since the new parameters may not be stochastic matrices (see Figure 6, for example the transition matrices $M_5^{-1}M_1$ or $M_5^{-1}M_2$ on the 2nd, 3rd and 4th trees depicted in the 13|24 leaf transformations might not be stochastic). In all these trees, the matrix attached to the interior edge remains untouched and is the matrix M_5 of the original tree.

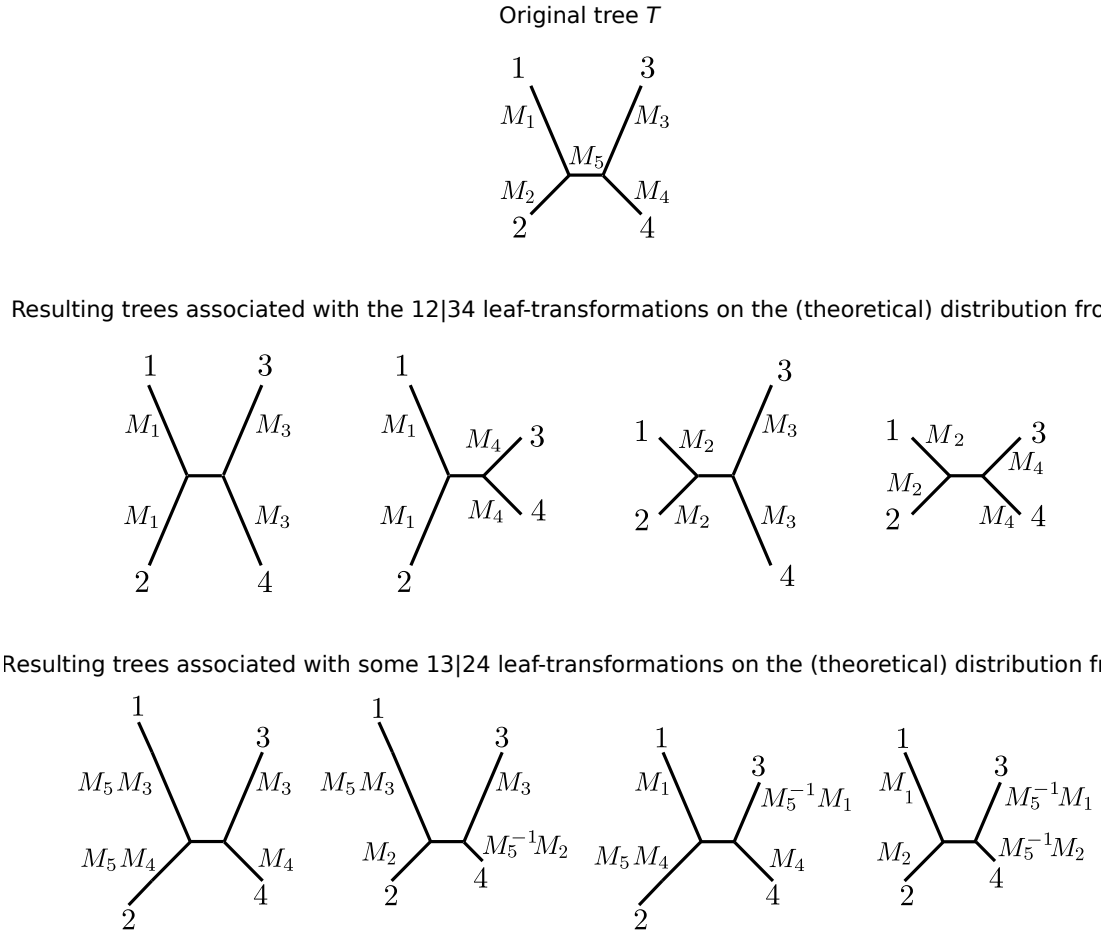

Figure 6: For the phylogenetic tree $T = 12|34$ (on the top of the figure), we show the effect on the parameters of applying the 12|34 leaf-transformations (middle) and some 13|24 leaf-transformations (bottom) on the theoretical distribution vector obtained from T.

The quotient score is well defined.— In this subsection we prove that if p arises from generic parameters on $T = 12|34$, the closest positive definite matrix to $F_{13|24}(\hat{p})$ has generically rank strictly larger than four, which is the last theoretical result needed jo justify the basis of `SAQ` method. For the sake of simplicity, the proof is carried out on the 3-parameter Kimura model. Then, a deformation argument can be applied to derive that the claim holds for generic Markov matrices.

Lemma 1. *Let p be the theoretical distribution from a 3-parameter Kimura process on the quartet tree $T = 12|34$ with generic transition matrices M_1, M_2, M_3, M_4 and M_5 (see the tree T of Figure 6). Then, if $\hat{p}$ is obtained by applying a $13|24$ leaf-transformation on p, it holds that*

$$\delta_4\left(psd\left(F_{13|24}(\hat{p})\right)\right) > 0.$$

Proof. We prove this result for a concrete $13|24$ leaf-transformation on p since the proof for the other transformations is analogous. Namely, we take the $13|24$ leaf-transformation that produces the vector $\hat{p}$ that corresponds to the tree $T = 12|34$ with transition matrix M_5M_3 at the edge adjacent to leaf 1, transition matrix M_5M_4 at the edge adjacent to leaf 2 and transition matrices M_3 and M_4 at the leaves adjacent to the leaves 3 and 4 respectively (see the first tree of the $13|24$ leaf-transformation shown in Figure 6).

It is known that for any matrix A, the rank of the nearest positive semidefinite matrix $psd(A)$ is equal to the number of positive eigenvalues of the symmetric matrix $\frac{1}{2}\left(A + A^t\right)$ (Higham 1988). So, it is enough to show that the symmetric matrix $B := \frac{1}{2}\left(F_{13|24}(\hat{p}) + F_{13|24}(\hat{p})^t\right)$ has at least 5 positive eigenvalues. To this aim, we use the following known result: if Δ_i is the i-th leading principal minor of a $n \times n$ matrix A, the number of sign changes in the sequence $\Delta_0 = 1, \Delta_1, \cdots, \Delta_n = det(A)$ equals the number of negative eigenvalues of the matrix A (see Ghys and Ranicki (2016) for a proof). Consequently, the number of positive eigenvalues of A equals the number of consecutive leading principal minors Δ_i without sign changes. We will also make use of Sylvester's *law of inertia* which states that any two $n \times n$ symmetric matrices X and Y have the same number of positive, negative and zero eigenvalues if and only if there exists an invertible matrix $H \in \mathbb{R}_{n\times n}$ such that $Y = H^tXH$ (see, for instance, Carrell (2017) for a precise statement and proof).

We start by writing the flattening matrix $F_{13|24}(\hat{p})$ in terms of these transition

matrices (Allman et al. 2012):

$$F_{13|24}(\hat{p}) = (M_5 M_3 \otimes M_3)^t \; D \; (M_5 M_4 \otimes M_4)$$

where $\otimes$ is the Kronecker product of matrices and D is the 16×16 diagonal matrix with the entries of $\frac{1}{4} M_5$. Consider the symmetric matrix

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

It is known that any 3-parameter Kimura matrix M diagonalizes through H. Moreover, the multiplicity of 1 as an eigenvalue of M is exactly one if M is generic (see Allman and Rhodes (2004)). Write $\{\alpha_k\}_{k=C,G,T}$ for the three eigenvalues of the matrix M_3 other than 1, and similarly $\{\beta_k\}_{k=C,G,T}$ and $\{\gamma_k\}_{k=C,G,T}$ for M_4 and M_5, respectively. We can assume that these eigenvalues are positive since the transition matrices should not be too far from the identity matrix.

Write $\bar{B} := (H \otimes H)^t B (H \otimes H)$. Because of Sylvester's law of inertia, both matrices, B and $\bar{B}$ have the same number of positive and negative eigenvalues. The matrix $\bar{B}$ can be written in terms of the eigenvalues of the transition matrices M_3, M_4 and M_5. Then, consider the 5×5 submatrix of $\bar{B}$ generated by removing the last 11 rows and columns of $\bar{B}$:

$$\bar{B}_5 = \frac{1}{4^4} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \alpha_C \beta_C & 0 & 0 & \alpha_C \beta_C \gamma_C^2 \\ 0 & 0 & \alpha_G \beta_G & 0 & 0 \\ 0 & 0 & 0 & \alpha_T \beta_T & 0 \\ 0 & \alpha_C \beta_C \gamma_C^2 & 0 & 0 & \alpha_C \beta_C \gamma_C^2 \end{pmatrix},$$

The leading principal minors of $\bar{B}_5$ are equal to the first five leading principal minors of $\bar{B}$:

$$\Delta_1 = 2, \quad \Delta_2 = 4\alpha_C \, \beta_C, \quad \Delta_3 = 8 \, \alpha_C \, \alpha_G \beta_C \, \beta_G, \quad \Delta_4 = 16 \, \alpha_C \, \alpha_G \, \alpha_T \, \beta_C \, \beta_G \, \beta_T,$$

$$\Delta_5 = -32 \, \alpha_C^2 \, \alpha_G \, \alpha_T \, \beta_C^2 \, \beta_G \, \beta_T \, (\gamma_C + 1) \, (\gamma_C - 1) \, \gamma_C^2.$$

We have that $\Delta_i > 0$ $(i = 1, 2, 3, 4)$ because α_k, β_k are positive for any $k \in \{C, G, T\}$. Moreover Δ_5 is also positive since $-(\gamma_C - 1) > 0$ for $0 < \gamma_C < 1$ and γ_G, γ_T are also positive. We conclude that the matrix $\bar{B}$ has at least 5 positive eigenvalues and so does B. This implies that $rk\left(psd\left(F_{13|24}(\hat{p})\right)\right)$ is greater than or equal to 5 and therefore, $\delta_4\left(psd\left(F_{13|24}(\hat{p}_i)\right)\right) > 0$. □

Random branch lengths

Here we present the ternary plots of the performance of `SAQ` when applied to the alignments generated on the tree 12|34 under the GM and homGTR models, with length 1 000 bp and 10 000 bp and random branch lengths uniformly distributed between 0 and 3.

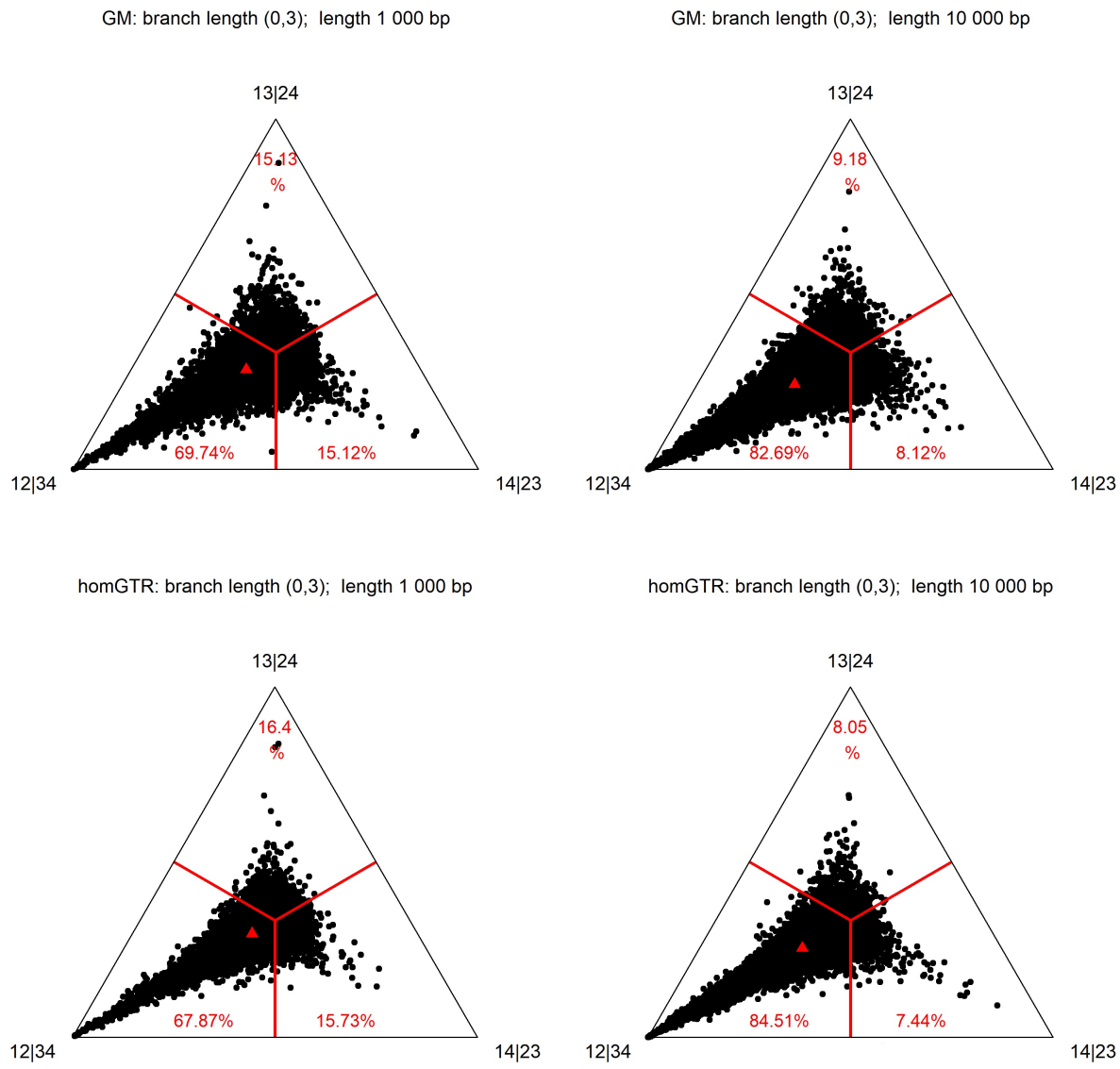

Figure 7: Ternary plots corresponding to the weights of `SAQ` applied to 10000 alignments generated under the 12|34 tree with random branch lengths uniformly distributed between 0 and 3. On each triangle the bottom-left vertex represents the underlying true topology 12|34, the bottom-right vertex is topology 13|24 and the top vertex is 14|23. Each triangle is divided into three regions according to which tree is selected by the method. The figures in red represent the percentage of alignments that correspond to the corresponding region according to `SAQ`. Top: correspond to data generated under GMM; bottom: data generated under homogenenous GTR. Left : 1000 bp; Right: 10000 bp.).