Journal of Theoretical Biology 440 (2018) 88–99

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Comparing evolutionary distances via adaptive distance functions

Yanir Damti[a], Ilan Gronau[b,*], Shlomo Moran[a], Irad Yavneh[a]

[a] Computer Science department, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel
[b] Efi Arazi School of Computer Science, The Herzliya Interdisciplinary Center (IDC), P.O.Box 167, Herzliya 46150, Israel

ARTICLE INFO

Article history:
Received 22 June 2017
Revised 19 December 2017
Accepted 21 December 2017
Available online 23 December 2017

Keywords:
Distance-based phylogenetic reconstruction
DNA substitution models
Adaptive distance estimation
Distance comparison
Fisher's linear discriminant

ABSTRACT

Distance-based methods for phylogenetic reconstruction are based on a two-step approach: first, pairwise distances are computed from DNA sequences associated with a given set of taxa, and then these distances are used to reconstruct the phylogenetic relationships between taxa. Because the estimated distances are based on finite sequences, they are inherently noisy, and this noise may result in reconstruction errors. Previous attempts to improve reconstruction accuracy focused either on improving the robustness of reconstruction algorithms to this stochastic noise, or on improving the accuracy of the distance estimates. Here, we aim to further improve reconstruction accuracy by utilizing the basic observation that reconstruction algorithms are based on a series of *comparisons* between distances (or linear combinations of distances). We start by examining the relationship between the stochastic noise in the sequence data and the accuracy of the comparisons between pairwise distance estimates. This examination results in improved methods for distance comparison, which are shown to be as accurate as likelihood-based methods, while being much simpler and more efficient to compute. We then extend these methods to improve reconstruction accuracy of quartet trees, and examine some of the challenges moving forward.

1. Introduction

Reliable estimation of evolutionary distances from molecular sequences is a core task in evolutionary inference, and particularly in phylogenetic reconstruction. Distance-based phylogenetic reconstruction is based on the following two-step approach: first, sequence data from n taxa are used to estimate the number of base substitutions between every pair of sequences in the evolutionary path that connects them; then, the $\binom{n}{2}$ pairwise distances are used to reconstruct the underlying phylogenetic tree. Distance-based algorithms for phylogenetic reconstruction are extremely popular due to their simplicity, reduced computational complexity, and theoretical guarantees. The premise behind this approach is that if the number of substitutions is accurately recovered in step 1 (up to a small tolerable error), then accurate reconstruction can be guaranteed in step 2 (Atteson, 1999). There have been numerous attempts to increase the robustness of distance-based reconstruction algorithms to noise in distances estimated from molecular sequence data (Erdos et al., 1999a,b; Gascuel, 1997; Gronau et al., 2012; Huson et al., 1999). On the other hand, there have been relatively few attempts to improve the accuracy of distance estimates. These attempts are based on controlling the complexity of the assumed substitution model (Zharkikh, 1994), or assuming that some parameters of the substitution process are shared among branches of the phylogeny (Hoyle and Higgs, 2003). However, until recently, the prevailing assumption has been that error in distance estimation is strictly determined by the assumed substitution model.

In a recent line of work, we showed that this basic assumption about noise in distance estimation is false (Gronau et al., 2009; 2010). We observed that while standard distance functions typically measure the total expected number of substitutions along an evolutionary path, weighted counts of different substitution types are also valid for the purpose of phylogenetic reconstruction. We suggested an adaptive approach for distance estimation, in which these weights are chosen to minimize the expected estimation error. While the theory behind this approach is relatively new, it has actually been used in practice in a few special cases. For example, in Kimura's 2 parameter model (Kimura, 1980), transition substitutions (A↔G and C↔T) are typically assumed to occur at a higher rate than transversion substitutions ({A, G}↔{C, T}). The standard distance function suggested for this model by Kimura (1980) estimates the total number of substitutions along a given evolutionary path. However, for very long evolutionary paths, in which saturation of transitions leads to noisy estimates of their counts, it is common practice to count only transversions using a formula from Cavender (1978); Farris (1973); Neymann (1971). Both formulas are

[*] Corresponding author.
E-mail address: ilan.gronau@idc.ac.il (I. Gronau).

https://doi.org/10.1016/j.jtbi.2017.12.022
0022-5193/© 2017 Elsevier Ltd. All rights reserved.

valid for the purpose of phylogenetic reconstruction because they estimate additive substitution counts under the K2P model (all substitutions versus only transversions). Consequently, every linear combination of these formulas is also additive, and may be considered as a valid distance function. Gronau et al. (2009) extended this basic observation to a wide range of substitution models, suggesting that distance functions should be *adapted* to the data being analyzed. The initial study focused on the estimation error of the length of a single evolutionary path, and Gronau et al. (2010) later demonstrated ways in which this approach can be utilized to improve the reconstruction accuracy for certain quartet trees.

In this paper we use the adaptive distance approach to formally examine the relationship between distance estimation noise and reconstruction accuracy. We do this by focusing on the fundamental problem of comparing two distances, which has been shown to provide a useful analytical framework for assessing reconstruction accuracy (Serdoz et al., 2017). This is because nearly all distance-based phylogenetic reconstruction algorithms can be broken up into a sequence of comparisons between linear combinations of distances. For instance, when reconstructing the phylogenetic tree over four taxa, *a, b, c, d*, the classical four-point method (FPM) has to determine which of the following distance sums is the smallest: $d(a,b)+d(c,d)$, $d(a,c)+d(b,d)$, or $d(a,d)+d(b,c)$ (Buneman, 1971). Another example is the Neighbor Joining (NJ) algorithm (Saitou and Nei, 1987), which iteratively joins the pair of taxa *i, j* that maximize $Q(i,j)=\sum_k d(i,k)+\sum_k d(j,k)-(n-2)d(i,j)$. An immediate consequence of the above observation is that the quality of distance based reconstruction depends on the accuracy of comparison queries, rather than on the numerical accuracies of the $\binom{n}{2}$ pairwise distances.

The main objective of this work is to harness the framework of adaptive distances to enhance the accuracy of comparison queries between distances or their linear combinations. We start by examining the simple problem of comparing the lengths of two independent evolutionary paths (Section 3). In that simple problem we use our notion of distance measures to achieve essentially the same accuracy as methods based on maximum-likelihood, but do that with much lower computational complexity. We then discuss extensions of this simple framework to the more general problem of comparing sums of distances and resolving quartet trees (Section 4).

2. Background

We start by presenting the required background on DNA substitution models, distance estimation via SR functions, and noise minimizing SR functions.

2.1. Kimura's two parameter substitution model

DNA sequence evolution is traditionally modeled using a process of base substitution along evolutionary paths connecting a set of taxa of interest. For a given path, this process is specified by a 4×4 *substitution rate matrix* **R** whose off-diagonal entries R_{ij} represent the (positive) rate of substitution from nucleotide *i* to nucleotide *j* along the path (diagonal elements are set to ensure rows sum up to 0). Here, we focus on Kimura's two parameter (K2P) model (Kimura, 1980), in which all transition substitutions (A↔G and C↔T) have the same rate (α) and all transversion substitutions ({A, G}↔{C, T}) have the same rate (β), which is typically smaller. A rate matrix in the K2P model is specified by the two parameters $\alpha,\ \beta>0$ and has the following form: (rows and columns are indexed by nucleotides A, G, C, and T, in that order)

$$\mathbf{R}(\alpha,\beta)=\begin{pmatrix}-\alpha-2\beta & \alpha & \beta & \beta\\ \alpha & -\alpha-2\beta & \beta & \beta\\ \beta & \beta & -\alpha-2\beta & \alpha\\ \beta & \beta & \alpha & -\alpha-2\beta\end{pmatrix}.$$

K2P is the simplest substitution model defined by more than one parameter, and a very commonly assumed model in molecular evolution. Another common assumption made when considering a collection of evolutionary paths is that they are *homogeneous*. In the K2P model this implies that all paths share the same transition-to-transversion (ti-tv) ratio $\kappa=\frac{\alpha}{2\beta}$. Thus paths in a homogeneous K2P model are often specified by the shared ti-tv ratio κ and the *total substitution rate* $t=\alpha+2\beta$ of every path. The substitution probabilities associated with an evolutionary path are given by exponentiating the rate matrix **R**. In K2P, this implies that the transition probability (p_{ti}) and the transversion probability (p_{tv}) are given by the following transformations:

$$p_{ti}=\frac{1}{4}\left(1+e^{-4\beta}-2e^{-2(\alpha+\beta)}\right);\quad p_{tv}=\frac{1}{4}\left(1-e^{-4\beta}\right). \tag{1}$$

The probability of observing an identical base on both sides of the path is given by $1-p_{ti}-2p_{tv}$.

2.2. Substitution rate (SR) functions

The length of an evolutionary path (distance between its end points) is defined by mapping the rate matrix **R** to a non-negative real value $d(\mathbf{R})\geq0$. Phylogenetic inference requires these distance measures to be *additive*, s.t. $d(\mathbf{R_1}+\mathbf{R_2})=d(\mathbf{R_1})+d(\mathbf{R_2})$. The standard additive measure used in most cases is the total substitution rate (t), which is simply the sum of all substitution rates (in K2P, $t=\alpha+2\beta$). In a previous study we showed that in a large class of substitution models, any positive linear combination of the eigenvalues of the rate matrix is additive (Gronau et al., 2009). We called these functions *substitution rate (SR) functions*, and examined ways of selecting SR functions for different phylogenetic inference tasks (see also Gronau et al., 2010). In K2P, the non-zero eigenvalues of a rate matrix $\mathbf{R}(\alpha,\ \beta)$ are -4β and $-2(\alpha+\beta)$, and thus every linear combination of α and β that is positive is an SR function (and all SR functions are of that form). This linear combination can be put in terms of the substitution probabilities p_{ti}, p_{tv} by exponentiating the rate eigenvalues:

$$\lambda\overset{\Delta}{=}e^{-4\beta}=1-4p_{tv};\quad \mu\overset{\Delta}{=}e^{-2(\alpha+\beta)}=1-2p_{ti}-2p_{tv}, \tag{2}$$

and representing the linear combination as follows:

$$-c_1\log(\lambda)-c_2\log(\mu)=(2c_2)\alpha+(4c_1+2c_2)\beta. \tag{3}$$

Because SR functions that are proportional to each other are equivalent, we may associate each SR function with the coefficient ratio $c=c_1/c_2$, and restrict our consideration to SR functions with $c_2=1$:

$$d_c(\lambda,\mu)=-c\log(\lambda)-\log(\mu)=2\alpha+(4c+2)\beta. \tag{4}$$

While valid SR functions might have a negative *SR coefficient c* (e.g. $d_{-\frac{1}{2}}(\lambda,\mu)=2\alpha$), we restrict our study here to non-negative SR coefficients. Thus, we consider SR functions ranging from $d_0(\lambda,\mu)=-\log(\mu)=2(\alpha+\beta)$ to $d_\infty(\lambda,\mu)=-\log(\lambda)=4\beta$ (the *transversion count* SR function). Another SR function of interest is $d_{\frac{1}{2}}$, which is proportional to the formula originally suggested by Kimura (1980) for the total rate:

$$\begin{aligned}d_{\frac{1}{2}}(\lambda,\mu)&=-\frac{1}{2}\log(\lambda)-\log(\mu)\\&=2\left(-\frac{1}{4}\log(1-4p_{tv})-\frac{1}{2}\log(1-2p_{ti}-2p_{tv})\right)=2t.\end{aligned} \tag{5}$$

2.3. Inference from observed data

In phylogenetic analysis, we wish to make inference from DNA sequences observed at the leaves of an evolutionary tree (phylogeny). Thus for each path connecting two leaves we are given a pairwise sequence alignment of length n, and we assume that every aligned pair of bases was independently generated from the same substitution process along the path. The likelihood of the pairwise alignment can then be expressed as a product of the appropriate substitution probabilities. In the K2P model, a pairwise alignment of length n with n_{ti} transition differences (e.g., [A,G]) and n_{tv} transversion differences (e.g., [A,T]) has the following likelihood:

$$\mathcal{L}(p_{ti}, p_{tv} | n, n_{ti}, n_{tv}) = p_{ti}^{n_{ti}} p_{tv}^{n_{tv}} (1 - p_{ti} - 2p_{tv})^{n - n_{ti} - n_{tv}}. \tag{6}$$

Maximum likelihood estimates (MLEs) of the model parameters can then be obtained by setting the substitution probabilities to values that maximize the likelihood function, and then applying the appropriate transformations (see Eq. (2)):

$$\hat{p}_{ti} = \frac{n_{ti}}{n}, \quad \hat{p}_{tv} = \frac{n_{tv}}{2n} \tag{7}$$

$$\hat{\lambda} = 1 - 4\hat{p}_{tv}, \quad \hat{\mu} = 1 - 2\hat{p}_{ti} - 2\hat{p}_{tv} \tag{8}$$

$$\hat{\alpha} = -\frac{1}{2}\log(\hat{\mu}) + \frac{1}{4}\log(\hat{\lambda}), \quad \hat{\beta} = -\frac{1}{4}\log(\hat{\lambda}). \tag{9}$$

If the transformation in Eq. (8) leads to negative values for $\hat{\lambda}$ or $\hat{\mu}$, then the path is considered to be saturated, and the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ are typically set to some arbitrary large value. For instance, if $\hat{\lambda}$ is negative, n_{tv} might be set to the largest possible value given the sequence length n, such that $\hat{\lambda}$ is positive (i.e., $n_{tv} = \lfloor \frac{n-1}{2} \rfloor$), and then, if $\hat{\mu}$ is negative n_{ti} can be set to the largest possible value such that $\hat{\mu}$ is positive (i.e., $n_{ti} = \lfloor \frac{n-1}{4} \rfloor$).

2.4. Distance estimation noise

Estimation of parameter values from data is an inherently noisy process. The noise, which is the difference between the true and inferred values, is determined by the parameter values and the length of pairwise alignment used in inference. This inherent noise transforms to distance estimation noise when applying an SR function to the estimated parameters. Gronau et al. (2009) showed that different SR functions have different noise patterns, and then demonstrated that this property can be used to improve the accuracy of phylogenetic reconstruction. These observations are based on two central derivations (see Gronau et al., 2009 for more details). The first is an expression approximating the mean square error (MSE) of a distance estimate obtained from $\hat{\lambda}, \hat{\mu}$ and an SR coefficient c:

$$\widetilde{MSE}(c; \hat{\lambda}, \hat{\mu}) = \frac{1}{n}\left(c^2\left(\frac{1}{\hat{\lambda}^2} - 1\right) + 2c\left(\frac{1}{\hat{\lambda}} - 1\right) + \frac{1}{2}\left(\frac{1}{\hat{\mu}^2} + \frac{\hat{\lambda}}{\hat{\mu}^2} - 2\right)\right). \tag{10}$$

The second is an expression for the SR coefficient that approximately minimizes the *relative distance estimation noise* defined by the normalized root MSE ($\sqrt{\widetilde{MSE}(c; \hat{\lambda}, \hat{\mu})}/d_c(\hat{\lambda}, \hat{\mu})$):

$$c^{opt}\left(\hat{\lambda}, \hat{\mu}\right) = \frac{\hat{\lambda}\left(\hat{\lambda}\log\hat{\lambda} + \hat{\lambda}^2\log\hat{\lambda} - 2\hat{\mu}^2\log\hat{\mu} - 2\hat{\lambda}\hat{\mu}^2\log\hat{\lambda} + 2\hat{\lambda}\hat{\mu}^2\log\hat{\mu}\right)}{2\hat{\mu}^2\left(1 - \hat{\lambda}\right)\left(\log\hat{\mu} - \hat{\lambda}\log\hat{\lambda} + \hat{\lambda}\log\hat{\mu}\right)}. \tag{11}$$

This noise-minimizing SR coefficient (c^{opt}) also plays a central role in our analysis here.

3. Comparing two paths

3.1. The two-path model

In this section we examine the fundamental task of inferring which of two independent evolutionary paths is longer. We assume that both paths obey the K2P substitution model and denote by (λ_1, μ_1) and (λ_2, μ_2) the exponentiated eigenvalues associated with the two K2P paths (see Eq. (2)). To avoid ambiguity as to which path is longer, we assume that the rates of both transitions and transversions are higher in one path than in the other. Specifically, we assume that path 1 is *inherently longer* than path 2, meaning that $\lambda_1 < \lambda_2$ and $\mu_1 < \mu_2$. We refer to two independent K2P paths satisfying these requirements as a *two-path model*, and our objective is to infer which path is longer by observing as input four DNA sequences of length n corresponding to the tips of the two paths.

Inference starts by examining the pairwise sequence alignment for each of the two paths and obtaining maximum likelihood estimates (MLEs) for the four model parameters: $\hat{\lambda}_1, \hat{\mu}_1, \hat{\lambda}_2, \hat{\mu}_2$. These estimates are obtained by a simple application of the MLE formula to each alignment separately (Eqs. (7) and (8)). What makes the inference problem complicated is that the estimated parameters might be *ambiguous* as to which path is longer. For instance, if $\hat{\lambda}_1 < \hat{\lambda}_2$ and $\hat{\mu}_1 > \hat{\mu}_2$, then inference will depend on the SR function we use to measure path length. In this section we examine this problem and suggest methods for selecting SR functions for this task.

3.2. Experiments on homogeneous models

We used experiments on simulated data to evaluate different inference methods for the two-path problem. In these simulations, we used homogeneous substitution models, which are commonly assumed in phylogenetic analysis. A homogeneous two-path model is defined by three parameters: the shared ti-tv ratio κ, and the total rates of the two paths t_1, t_2. In our simulations path 1 is longer than path 2, hence $t_1 > t_2$. We simulated the substitution process on this homogeneous model using sequences of length $n = 500$ bp. In cases where a simulated path is saturated, we adopted the strategy described in the end of Section 2.3. We then applied a series of inference methods to the four sequences and recorded whether the method successfully inferred that path 1 is longer than path 2. We repeated this process using 100,000 independent simulations to obtain an estimated success rate for each tested inference method on the two-path model in question. By using 100,000 independent simulations, an estimate of p for the success ratio has a standard error that is approximately $\sqrt{p(1-p)/100,000}$. This standard error is bounded from above by 0.16% (maximum obtained at $p = 0.5$).

One of the benefits of using homogeneous models in our experiments is that homogeneous models have a statistically optimal, yet computationally intensive, inference method. Notice that if the estimated model parameters $\hat{\lambda}_1, \hat{\mu}_1, \hat{\lambda}_2, \hat{\mu}_2$ are consistent with a homogeneous model, then they are unambiguous as to which path is longer, because $\frac{\log(\hat{\mu}_1)}{\log(\hat{\lambda}_1)} = \frac{\log(\hat{\mu}_2)}{\log(\hat{\lambda}_2)}$ implies that $\{\hat{\lambda}_1, \hat{\lambda}_2\}$ and $\{\hat{\mu}_1, \hat{\mu}_2\}$ are ordered the same way. This ordering directly determines the inference outcome regardless of the chosen SR function. The problem with this approach is that MLEs under a homogeneous restriction no longer have closed-form solutions and they require applying computationally intensive optimization techniques. This is because the ti-tv ratio κ is shared between the two likelihoods of the two alignments that make up the data. Nonetheless,

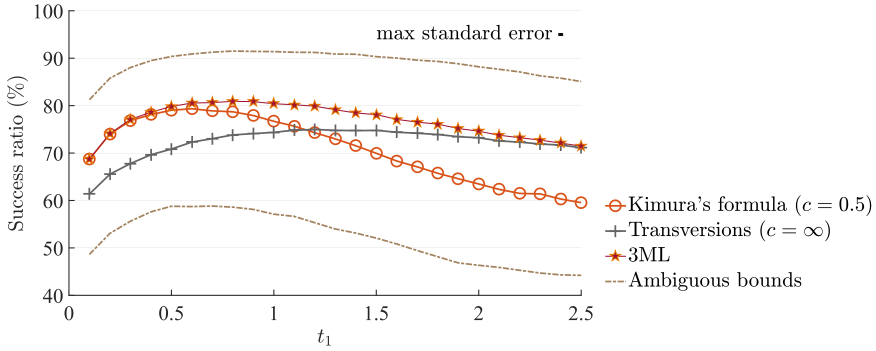

Fig. 1. Basic methods for inference of longest path applied to homogeneous two-path models with $\kappa = 2$ and $t_2/t_1 = 0.9$. Success ratios were estimated for the two standard SR functions (Kimura's formula and the transversion only formula) and the 3ML method. Notice that the success ratios of all methods fall within the ambiguous range bounded from below by the number of unambiguous correct cases (lower dashed line) and from above by the number of cases that are not unambiguously incorrect (upper dashed line). Results are based on 100,000 independent experiments run for each model using 500 bp long sequences. Standard errors for success ratio estimates are less than 0.16% (See Section 3.2).

MLEs for the three free parameters of a homogeneous two-path model ($\hat{\kappa}, \hat{t}_1, \hat{t}_2$) can be inferred by numeric optimization, and the path inferred to be longer is the one whose total rate ($\hat{t}_i$) is estimated to be larger. We refer to this method of inference as the *3ML method* and use it as a point of comparison for the methods we propose. Our objective is thus to propose inference methods that: (1) do not assume that the model is homogeneous, (2) are much more efficient than 3ML, and (3) whose accuracy is comparable to that of the 3ML method in homogeneous models.

3.3. Resolving ambiguity

Consider a two-path model with path 1 inherently longer than path 2, meaning that $\lambda_1 < \lambda_2$ and $\mu_1 < \mu_2$. We obtain the four parameter MLEs $\hat{\lambda}_1, \hat{\mu}_1, \hat{\lambda}_2, \hat{\mu}_2$ and are interested in using them to determine which path is longer. This is done by choosing an SR coefficient ($c \geq 0$) and comparing the two distances $d_c(\hat{\lambda}_i, \hat{\mu}_i) = -c\log(\hat{\lambda}_i) - \log(\hat{\mu}_i)$ for $i \in \{1, 2\}$. Notice that in some cases, the result of this comparison depends on the choice of c, whereas in others it does not. For instance, if $\hat{\lambda}_1 < \hat{\lambda}_2$ and $\hat{\mu}_1 < \hat{\mu}_2$, then for all $c \geq 0$ we have $d_c(\hat{\lambda}_1, \hat{\mu}_1) > d_c(\hat{\lambda}_2, \hat{\mu}_2)$ and path 1 is always inferred to be longer. We refer to such cases as being *unambiguously correct*. We similarly define *unambiguously incorrect* cases as ones in which $\hat{\lambda}_1 > \hat{\lambda}_2$ and $\hat{\mu}_1 > \hat{\mu}_2$, and refer to the remaining cases, where $\{\hat{\lambda}_1, \hat{\lambda}_2\}$ and $\{\hat{\mu}_1, \hat{\mu}_2\}$ are ordered differently, as *ambiguous*. Our study naturally focuses on ambiguous cases, where the path inferred to be longer depends on the choice of SR function. To better understand ambiguity, we examine the difference between the two inferred path lengths as a function of c:

$$\begin{aligned}\Delta(c; \hat{\lambda}_1, \hat{\mu}_1, \hat{\lambda}_2, \hat{\mu}_2) &= d_c\left(\hat{\lambda}_1, \hat{\mu}_1\right) - d_c\left(\hat{\lambda}_2, \hat{\mu}_2\right) \\ &= \log\left(\frac{\hat{\lambda}_2}{\hat{\lambda}_1}\right)c - \log\left(\frac{\hat{\mu}_1}{\hat{\mu}_2}\right). \qquad (12)\end{aligned}$$

Because this difference is a linear function of c, then it has a single X-intercept at $c_{switch} = \log(\frac{\hat{\mu}_1}{\hat{\mu}_2})/\log(\frac{\hat{\lambda}_2}{\hat{\lambda}_1})$. SR functions on one side of the intercept result in correct inference and SR functions on the other side result in incorrect inference. In ambiguous cases one side is represented by $d_0(\lambda, \mu) = -\log(\mu)$ and the other is represented by the transversion count $d_\infty(\lambda, \mu) = -\log(\lambda)$. Our objective is thus to develop methods for determining which 'side' of c_{switch} is more likely to be correct.

We ran a series of experiments to examine the effect of ambiguity on the inference task. We considered a series of homogeneous two-path models with $\kappa = 2$ and ${}^{t_2}/_{t_1} = 0.9$, and recorded the number of data sets that fall in each of the three types. In addition, we recorded the accuracy of two standard SR functions: Kimura's formula ($d_{\frac{1}{2}}$) and the transversion only formula (d_∞) as well as the 3ML method. All models considered in these simulations resulted in 31%–42% ambiguous cases (Fig. 1; size of interval between the two dashed lines). As expected, the success ratio of all methods was bounded from below by the number of unambiguous correct cases and bounded from above by the number of cases that were not unambiguous incorrect (dashed lines in Fig. 1). As observed in previous studies for other phylogenetic inference tasks, Kimura's SR function performs well when the paths are short, the transversion-based distance performs well for long paths (Gronau et al., 2009), and the ML-based approach provides an upper bound in terms of accuracy.

3.4. Choosing a discriminating SR function

One way to address the issue of ambiguity is to measure how much an SR function *discriminates* between the two path lengths. The measure we use for discrimination is based on Fisher's linear discriminant (Fisher, 1936) and takes into consideration the difference between the two distances and the magnitude of noise involved in their estimation (see Eqs. (12) and (10)).

$$DiscScore(c) = \frac{\left(\Delta(c; \hat{\lambda}_1, \hat{\mu}_1, \hat{\lambda}_2, \hat{\mu}_2)\right)^2}{\widetilde{MSE}(c; \hat{\lambda}_1, \hat{\mu}_1) + \widetilde{MSE}(c; \hat{\lambda}_2, \hat{\mu}_2)}. \qquad (13)$$

An SR function that maximizes this *discrimination score* is expected to result in a large difference between estimated path lengths relative to the distance estimation noise, and is thus relatively likely to lead to accurate inference. We tested this assertion on the simulated data from Fig. 1 and saw that the success rates of this inference method (select the function which maximizes *DiscScore*) were typically higher than those of the standard SR functions, but somewhat lower than the success rates of the 3ML method (Fig. 2). We then tested what would happen if we infer the shorter path by an SR function maximizing the discrimination score computed with the *true model parameters* λ_i, μ_i. While this approach does not constitute a valid inference method because it makes use of the unknown true parameters values, we view it as an 'oracle' suggesting an SR function for a given two-path model. As such, it provides insight into what makes a good SR function. We see that for nearly all models, the SR function chosen by *DiscScore*-oracle has a success rate that is practically identical to that of the 3ML method. This fact was further confirmed by checking that for each model examined in these simulations, no single SR function had a success rate that was significantly higher. Thus

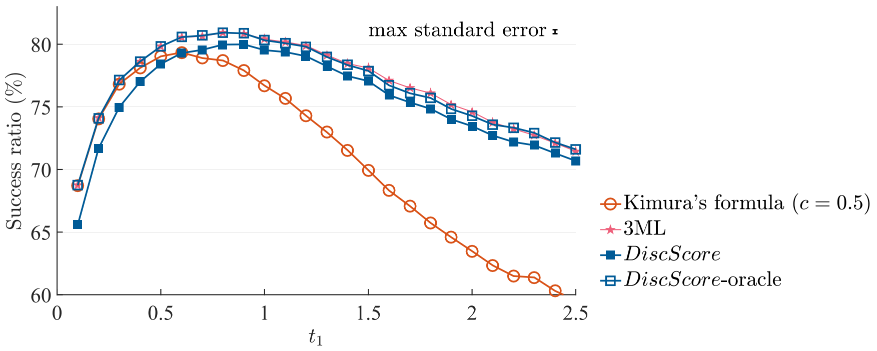

Fig. 2. Inference methods based on Fisher's linear discriminant applied to homogeneous two-path models with $\kappa = 2$ and $t_2/t_1 = 0.9$. Success ratios were estimated for methods based on SR functions maximizing the discrimination score (Eq. (13)) computed from the inferred parameter values (*DiscScore*) and the (unobserved) true parameter values (*DiscScore*–oracle). Success ratios for Kimura's formula and the 3ML method are shown for reference. The success ratios of *DiscScore*–oracle are near optimal, and the success ratios of *DiscScore* are somewhat lower, likely due to the influence of noise (see Section 3.5). The simulated data sets are the ones generated for Fig. 1, and the maximum standard error of a success ratio estimate is 0.16%.

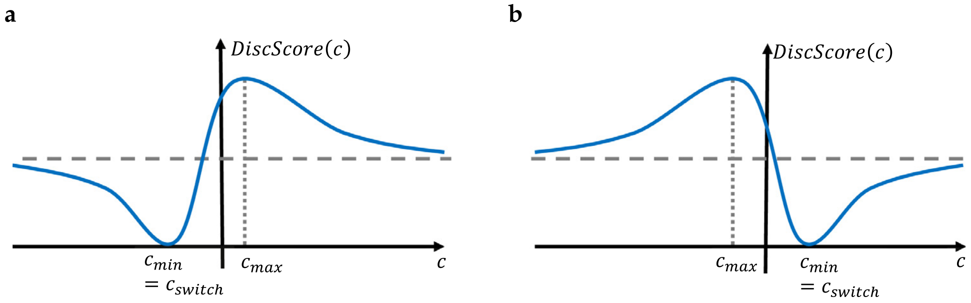

Fig. 3. *DiscScore*(c) is a rational function with quadratic polynomials in the numerator and in the denominator. It has two extrema: one maximum ($c_{\max}$) and one minimum at c_{switch} (*DiscScore*(c_{switch}) $= 0$). *DiscScore*(c) approaches the same positive value at $\pm\infty$. The graphs here depict two forms of *DiscScore*(c) observed in our experiments. **(a)** In a typical unambiguous case ($c_{switch} < 0$), the global maximum is obtained at a positive point $c_{\max} > 0$, and $d_{c_{\max}}$ is selected by the *DiscScore* method. **(b)** In a typical ambiguous case ($c_{switch} > 0$), the global maximum is negative $c_{\max} < 0$, which implies that *DiscScore* selects as an SR function either d_0 or d_∞.

Fisher's linear discriminant provides an excellent oracle for selecting an SR function for a given two-path model, but it appears to be less effective when applied as an inference method to the observed data.

3.5. The noise-minimizing SR functions c_1^{opt} and c_2^{opt}

We examined the behavior of Fisher's linear discriminant more closely to understand the discrepancy between the *DiscScore* inference method based on estimated parameter values and *DiscScore*-oracle based on true values used in simulation. Along the entire real axis ($c \in [-\infty, +\infty]$), the discrimination score typically has a single local maximum and a single local minimum. Since the discrimination score is non-negative, the local minimum is attained at c_{switch}, where the score is zero. We noticed that in all models considered, the maximum point of the discrimination score computed from the true parameters was at a positive and finite SR coefficient ($0 < c < \infty$), which was thus the one used by *DiscScore*-oracle. This was also true for the discrimination score computed from the estimated parameters in the great majority of unambiguous cases (Fig. 3(a)). However, in *all* ambiguous cases encountered in our experiments, the local (and global) maximum point was attained in a negative SR coefficient (Fig. 3(b)). Thus in every ambiguous case, the non-negative SR coefficient that maximized the discrimination score was either $c = 0$ ($d_c(\lambda, \mu) = -\log(\mu)$) or $c = \infty$ ($d_c(\lambda, \mu) = -\log(\lambda)$). This means that the *DiscScore* method is practically the same as choosing the SR function that maximizes the discrimination score among the two extreme SR functions d_0 and d_∞.

This simple observation implies that the *DiscScore* method can be implemented efficiently by two simple computations rather than having to find the maximum in the entire range. However, more importantly, it provides some insight into why this method is not as good as the oracle based on the same score. Because they lie on the edges of the range of SR functions, d_0 and d_∞ have large relative distance estimation noise, and the discrimination score could be a poor indicator for their expected accuracy. False inference might be a result of a deceptively high discrimination score for one of these SR functions.

We chose to address this problem by using less noisy SR functions as candidates for inference. Consider the two SR functions that minimize the relative distance estimation noise for the two paths in the model. The coefficients corresponding to these *noise-minimizing* SR functions are approximated by Eq. (11): $c_i^{opt} \triangleq c^{opt}(\hat{\lambda}_i, \hat{\mu}_i)$. Both SR coefficients are associated with low levels of distance estimation noise for the two paths, and so are other SR coefficients in the range between them, which we refer to as the c^{opt} *range*. Consequently, we define c^{opt}-*ambiguous* cases as data sets in which c_{switch} falls in the c^{opt} range. We examined the influence of this focus on the c^{opt} range on our inference experiments. First, the number of c^{opt}-ambiguous cases was much lower than the number of overall ambiguous cases, and did not exceed 20% (Fig. 4). Importantly, while the number of unambiguous correct cases increased by 17%–21%, the number of unambiguous incorrect cases increased by only 6%–13%, resulting in overall improvement in inference accuracy. Consequently, the simple method that infers the long path based on an SR coefficient selected uniformly at random from $\{c_1^{opt}, c_2^{opt}\}$ is only slightly less accurate than the 3ML method and comparable to the *DiscScore* method proposed in the previous section.

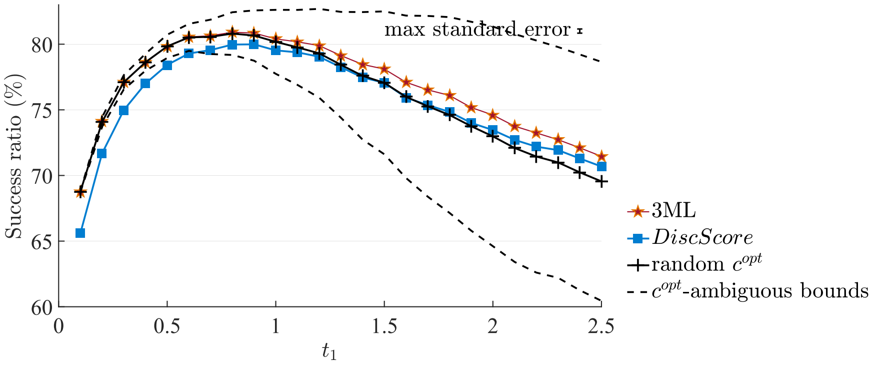

Fig. 4. Basic inference methods based on the noise-minimizing SR coefficients c_1^{opt} and c_2^{opt} applied to homogeneous two-path models with $\kappa = 2$ and $t_2/t_1 = 0.9$. Dashed lines represent proportions of c^{opt}-unambiguous correct cases (bottom) and cases that are not c^{opt}-unambiguous incorrect (top). Success ratios for *DiscScore* and the 3ML method are shown for reference. A simple inference method based on randomly choosing SR coefficients among c_1^{opt} and c_2^{opt} (random c^{opt}) is only slightly less accurate than the 3ML method, and is comparable to *DiscScore*. The simulated data sets are the ones generated for Fig. 1, and the maximum standard error of a success ratio estimate is 0.16%.

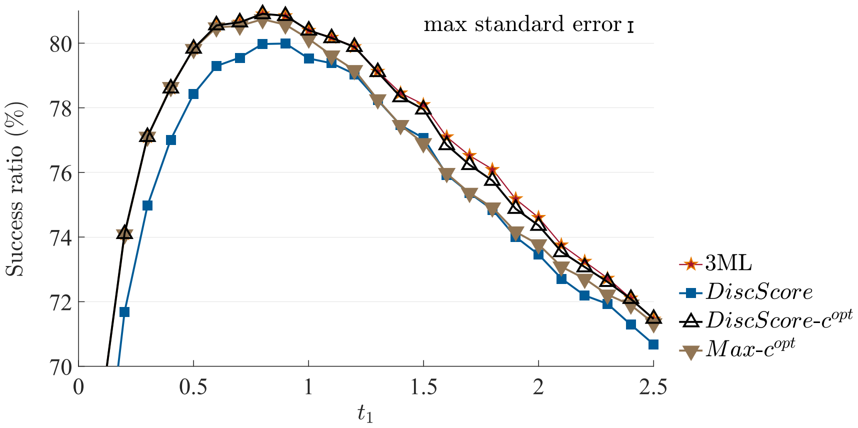

Fig. 5. Advanced inference methods based on the noise-minimizing SR coefficients c_1^{opt} and c_2^{opt} applied to homogeneous two-path models with $\kappa = 2$ and $t_2/t_1 = 0.9$. We considered different scores for evaluating the two noise-minimizing SR functions: the discrimination score (*DiscScore-c^{opt}*), and the maximum coefficient (*Max-c^{opt}*). The *DiscScore-c^{opt}* method has near optimal accuracy, and the simpler *Max-c^{opt}* method has slightly lower success rates, especially for models with intermediate path lengths. The simulated data sets are the ones generated for Fig. 1, and the maximum standard error of a success ratio estimate is 0.16%. Results are shown for models with $t_1 \geq 0.2$ to enable focus on subtle differences between methods.

3.6. Selecting between c_1^{opt} and c_2^{opt}

Next, we examined how an informed choice between the two noise-minimizing SR functions would improve inference accuracy compared to the random choice we examined in Fig. 4. One way to select between the two SR functions is by using the discrimination score we defined in Section 3.4. We tested this method, which we termed *DiscScore-c^{opt}*, on our simulated data and saw that it had a considerably higher success rate than the original *DiscScore*, which maximizes the discrimination score across the entire range of positive SR coefficients (Fig. 5). Furthermore, *DiscScore-c^{opt}* had success rates that were statistically identical to those of the 3ML method across nearly all models tested. We attribute the improved performance to the fact that we are using the discrimination score to evaluate SR functions with low distance estimation noise (see discussion in Section 3.5). We tested this approach also in models with various ti-tv ratios and various ratios between path lengths $^{t_2}/_{t_1}$ (Supplementary Figure S1), and observed success rates that were very similar to those of the 3ML method in all tested models. Importantly, this near-optimal accuracy is achieved by an inference method that is much simpler than the 3ML method and involves the following steps:

1. Compute the two noise-minimizing SR coefficients c_1^{opt} and c_2^{opt} using Eq. (11).
2. If the path that is longer under SR function $d_{c_1^{opt}}$ is also longer under $d_{c_2^{opt}}$, then return the identity of this path.
3. Otherwise, compute the discrimination scores of c_1^{opt} and c_2^{opt} using Eq. (13) and return the identity of the path that is longer based on the SR coefficient with higher score.

This approach is appealing because it is very efficient and does not require numeric maximization of a complex likelihood function. Indeed, our experiments indicate that *DiscScore-c^{opt}* is roughly 46 times faster than 3ML (Supplementary Table S1). Another advantage of this approach is that in cases that are tougher for inference, where the two paths have similar lengths, the two SR coefficients c_1^{opt} and c_2^{opt} are also very similar and the third step is not reached. We considered several selection criteria as alternatives to the discrimination score for step 3, (e.g., ratio between inferred lengths, sum of normalized MSE, distance from c_{switch}), and they all had slightly lower success rates. One alternative method worth noting is based on selection of the *larger SR coefficient* among c_1^{opt} and c_2^{opt}. The rationale behind this method, which we call *Max-c^{opt}*, is that two-path models where c^{opt}-ambiguous cases are prevalent typically have at least one long path (see Fig. 4), and in these models large SR coefficients typically have high success rates (see d_∞ in Fig. 1). As expected, this method has near optimal success rates for models with short paths (due to few ambiguous cases) and models with long paths (due to choice of large SR coefficient), and some-

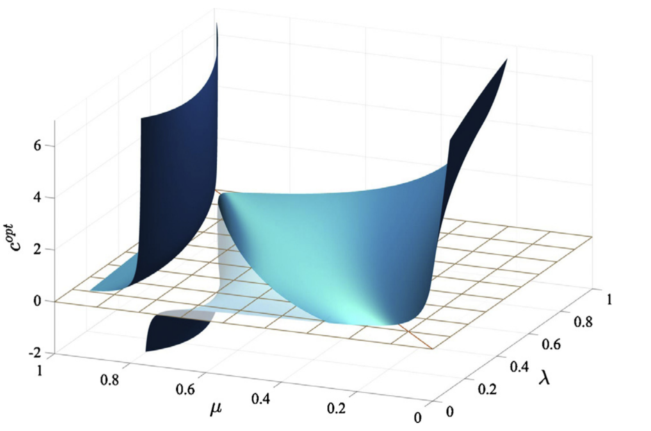

Fig. 6. The 2D function $c^{opt}(\lambda, \mu)$ defined by Eq. (11) computed in the range $[0, 1]^2$ in a grid with 0.01 intervals. In this range, we see that c^{opt} is increasing with λ and decreasing with μ, establishing Claim 1. The biologically-relevant subregion where $\lambda > \mu$ is shown to the right of the diagonal red line. The discontinuation line between the two manifolds corresponds to points where the denominator of Eq. (11) is zero.

what lower relative success rates in models with paths of intermediate length (Fig. 5). We revisit this method later in Section 4 when we extend the comparison problem to sums of distances.

3.7. Self contraction of noise-minimizing SR function

We encountered an interesting observation about c^{opt}-ambiguous cases in our experiments (Fig. 4). In *all* these cases, the SR function that correctly inferred that path 1 was longer than path 2 was the one corresponding to coefficient c_2^{opt}. In other words, if the two SR functions minimizing the distance estimation noise for the two paths do not agree as to which path is longer, then the SR function that minimizes the distance estimation noise of *the shorter path* is the one that results in the correct answer. This observation is a consequence of a property of noise-minimizing SR functions, which we call *self contraction*:

Lemma 1 (self contraction)**.** *Let $\hat{\lambda}_1, \hat{\mu}_1, \hat{\lambda}_2, \hat{\mu}_2$ be estimates of the four parameters in a two-path model, and let c_1^{opt} and c_2^{opt} be the two noise-minimizing SR functions defined by Eq. (11): $c_i^{opt} = c^{opt}(\hat{\lambda}_i, \hat{\mu}_i)$. If c_{switch} falls between these two coefficients (c^{opt}-ambiguous case), then for every $i \in \{1, 2\}$, the SR function corresponding to c_i^{opt} infers that path i is the shorter path among the two: $d_{c_i^{opt}}(\hat{\lambda}_i, \hat{\mu}_i) < d_{c_i^{opt}}(\hat{\lambda}_{3-i}, \hat{\mu}_{3-i})$.*

Note that this lemma implies that if path 1 is longer, then c_2^{opt} will lead to correct inference in all c^{opt}-ambiguous cases, as we observed in our simulations. This self-contraction property follows from a basic property of the formula for the noise-minimizing SR coefficient, which we confirmed by computing the values of $c^{opt}(\lambda, \mu)$ in the range $[0, 1]^2$ (see Fig. 6):

Claim 1. *Let $c^{opt}(\lambda, \mu)$ denote the noise-minimizing SR coefficient defined by Eq. (11). Then c^{opt} is a monotonic increasing function of λ and a monotonic decreasing function of μ.*

Proof of Lemma 1. (self contraction). For a given SR coefficient, c, denote by $\Delta(c) = d_c(\hat{\lambda}_1, \hat{\mu}_1) - d_c(\hat{\lambda}_2, \hat{\mu}_2)$ the difference between the lengths of the two paths according to c (see Eq. (12)). We need to prove that $\Delta(c_2^{opt}) > 0$ and $\Delta(c_1^{opt}) < 0$. Because we assume a c^{opt}-ambiguous case, then one of these differences is positive and the other is negative. We can thus prove that $\Delta(c_2^{opt}) > 0$ by showing that $\Delta(c_2^{opt}) > \Delta(c_1^{opt})$. Recall that in ambiguous cases, $\{\hat{\lambda}_1, \hat{\lambda}_2\}$ and $\{\hat{\mu}_1, \hat{\mu}_2\}$ are ordered differently. If we assume $\hat{\lambda}_1 < \hat{\lambda}_2$, then we have $\hat{\mu}_1 > \hat{\mu}_2$, implying through Claim 1 that $c_1^{opt} < c_2^{opt}$, and through Eq. (12) that $\Delta(c)$ is a monotonic *increasing* function in c. Thus, $\Delta(c_1^{opt}) < \Delta(c_2^{opt})$, as required. If, on the other hand, $\hat{\lambda}_1 > \hat{\lambda}_2$, then $\hat{\mu}_1 < \hat{\mu}_2$, implying through Claim 1 that $c_2^{opt} < c_1^{opt}$ and through Eq. (12) that $\Delta(c)$ is a monotonic *decreasing function*. Thus we still have $\Delta(c_1^{opt}) < \Delta(c_2^{opt})$, as required. □

The self contraction property is nicely demonstrated in Fig. 7, which depicts data simulated under models with identical paths. We generated data for a series of homogeneous two-path models with $t_1 = t_2$ and $\kappa = 2$ and measured the inference bias of different methods. As expected, all inference methods that are solely based on the observed alignments (Kimura's formula, 3ML, and *Max-c^{opt}*) show no bias and infer both paths as being longer roughly the same number of times. However, a method based on the SR function that minimizes the distance estimation noise of path 2 shows significant inference bias toward path 1.

In principle, we would like to use this bias caused by the self-contraction property to improve inference accuracy. Note that by using the SR function that minimizes the noise for the shorter of the two paths we could obtain success rates that are even higher than those of the 3ML method (upper dashed line in Fig. 4). However, because this requires knowing which path is the shorter one, then this property does not end up having practical implications when comparing two paths. Nonetheless, it may have interesting implications in the context of phylogenetic inference (see Discussion section).

4. Resolving quartets and comparing sums of path lengths

4.1. Quartet reconstruction

In this section we examine ways of extending methods presented in Section 3 for the problem of comparing two paths to the problem of phylogenetic reconstruction. We focus on quartet trees to demonstrate the usefulness of these methods and outline potential challenges. A quartet is a phylogenetic tree representing the evolution of four taxa (associated with leaves of the tree). In its unrooted form, a quartet has four external edges (edges that touch leaves) and a single internal edge (Fig. 8). The objective of phylogenetic reconstruction in the case of a quartet is to infer the topology of the tree, which is typically represented by the split that the internal edge induces on the leaves. For instance, split (*a, b*|*c, d*) means that all four paths connecting taxa *a, b* with taxa *c, d* traverse through the internal edge.

The split of a quartet can be inferred quite simply by observing the lengths of the six paths connecting all pairs of taxa. Denote by $d(i, j)$ the length of the path connecting taxa *i* and *j*. Since a path length is the sum of lengths of edges that make up that path, we get for a quartet with split (*a, b*|*c, d*):

$$d(a, c) + d(b, d) = d(a, d) + d(b, c) > d(a, b) + d(c, d). \qquad (14)$$

The difference between the first two sums in Eq. (14) and the third one equals twice the length of the internal edge. Hence, the quartet split can be inferred by examining the three sums of path lengths specified in Eq. (14), and choosing the split corresponding to the smallest sum (the two paths considered in the sum do not contain the internal edge). This inference method is called the four point method (FPM; see Buneman, 1971 and Sattath and Tversky, 1977), and it is the cornerstone of distance-based phylogenetic reconstruction methods.

Inference of quartets using the FPM depends on our ability to accurately compare sums of distances, thus the approaches we introduced in Section 3 to compare path lengths may be useful in this setting as well. However, there are some potential complications that stem from differences between the two-path prob-

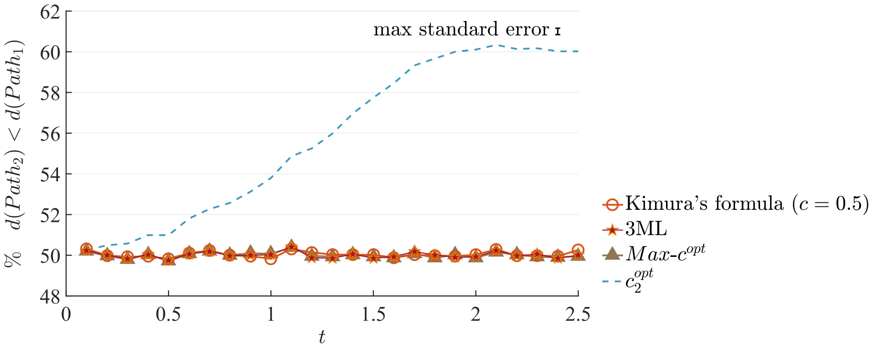

Fig. 7. Distinguishing between two identical paths in homogeneous two-path models with $\kappa = 2$ and $t_2 = t_1$. We considered three different inference methods (see legend) as well as a method that uses the SR function that minimizes the distance estimation noise of path 2 (c_2^{opt}). For each method we measured the proportion of times in which it infers that path 1 is longer than path 2, which we expect to be 50%. Indeed, the three inference methods are all observed to be unbiased, but the method based on c_2^{opt} infers path 1 to be longer in as much as 60% of the data sets for models with long paths. This biased inference is a direct result of the self-contraction property (Lemma 1). Results are based on 100,000 independent experiments run for each model using 500 bp long sequences. Standard errors for success ratio estimates are less than 0.16% (See Section 3.2).

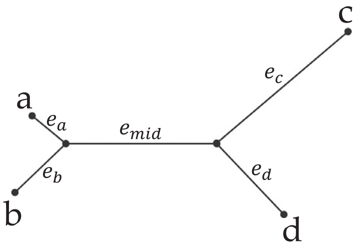

Fig. 8. A general diagram of an unrooted quartet. The quartet has four external edges and a single internal edge. The length of each edge corresponds to the evolutionary distance between the two nodes it touches. The split induced by this quartet is designated by (a, b|c, d).

lem and quartet reconstruction: (1) the FPM compares sums of lengths and not the lengths directly, (2) the inference task has three possible outcomes and not two, and (3) the six paths considered share common edges and are thus not independent as assumed in the two-path model. In the remainder of this section we suggest extensions of the adaptive distance methods presented in Section 3 to quartet inference and we examine the impact of these three issues.

4.2. Ambiguity in quartet reconstruction

Applying the FPM to data, we first have to obtain length estimates for the six paths in Eq. (14). This is done by estimating MLEs for the K2P eigenvalues $\hat{\lambda}_{ij}, \hat{\mu}_{ij}$ for each of the six paths connecting pairs of taxa i, j, and applying an SR function to these estimates. The additivity of SR functions (see Section 2.2) guarantees that the length of each path equals the sum of lengths of the edges that make up that path. Let us denote the 12 parameter estimates by $(\hat{\Lambda}, \hat{M})$, and for a given SR coefficient c and a given pair of taxa $\{i, j\} \subset \{a, b, c, d\}$, denote by $d_c(i, j) = d_c(\hat{\lambda}_{ij}, \hat{\mu}_{ij})$. The FPM is implemented by computing the following 3-dimensional vector:

$$FPMsums\left(c; \hat{\Lambda}, \hat{M}\right) = (d_c(a,b) + d_c(c,d), d_c(a,c) + d_c(b,d), d_c(a,d) + d_c(b,c)). \tag{15}$$

The choice of SR function influences the outcome of the FPM by affecting the entries of the *FPMsums* vector. Recall that in the two-path problem, inference was based on determining the smallest entry in a two-dimensional vector $(d_c(\hat{\lambda}_1, \hat{\mu}_1), d_c(\hat{\lambda}_2, \hat{\mu}_2))$. We observed that entries in this two-dimensional vector were linear functions of the SR coefficient c, and ambiguous inference was a result of the intersection of these two lines at a positive SR coefficient c_{switch}. In the quartet reconstruction problem, the three entries of $FPMsums(c; \hat{\Lambda}, \hat{M})$ are also linear functions of the SR coefficient c (e.g., $d_c(a,b) + d_c(c,d) = -\log(\hat{\lambda}_{ab}\hat{\lambda}_{cd})c - \log(\hat{\mu}_{ab}\hat{\mu}_{cd})$). Ambiguous quartet inference is thus a result of intersection points between these lines that satisfy the following conditions (Fig. 9): (1) the intersection is at a positive SR coefficient $c_{switch} > 0$, and (2) none of the three lines passes below the intersection point at c_{switch}. Because you cannot have three intersection points that satisfy the second condition, there are at most two switch points for inference. Furthermore, inference is *convex* in the sense that if SR coefficients c_1 and c_2 result in the same inferred split, then so does any SR coefficient in the range between them. This implies that ambiguity in quartet inference is similar to ambiguity in the two-path problem described in Section 3, with the main difference being that there are possibly three results for inference instead of two.

4.3. Adaptive quartet inference methods

One of the main conclusions from examining the two-path problem was that it is very useful to restrict consideration to SR coefficients which are likely to be less noisy. In the quartet model we have six paths, each of which has its noise-minimizing SR coefficient $c_{i,j}^{opt} = c^{opt}(\hat{\lambda}_{ij}, \hat{\mu}_{ij})$ (Eq. (11)). The c^{opt}-range is thus defined by the minimal range containing these six noise-minimizing SR coefficients. If all six SR functions agree on the inferred split (the smallest entry in the *FPMsums* vector is in the same index), then this split is returned. If they disagree, then we use some criterion to decide which SR function is more likely to be correct.

A criterion that proved to be very useful in the two-path model was the discrimination score based on Fisher's linear discriminant (Eq. (13)). Extending this score to the quartet model requires considering all different noise estimates in the denominator and the relevant difference in the numerator (squared). In the two-path model, this was simply the difference between the two path lengths. Here, we consider the difference between the smallest entry in the *FPMsums* vector and the second smallest entry. We denote this difference by $\Delta FPMsums$, and define the quartet discrimination score as follows:

$$DiscScoreQuart\left(c; \hat{\Lambda}, \hat{M}\right) = \frac{\left(\Delta FPMsums\left(c; \hat{\Lambda}, \hat{M}\right)\right)^2}{\sum_{\{i,j\} \subset \{a,b,c,d\}} \widetilde{MSE}(c; \hat{\lambda}_{ij}, \hat{\mu}_{ij})} \tag{16}$$

The *DiscScoreQuart* method for inferring a quartet split is thus given by applying the FPM on distances obtained by using the SR function that maximizes the quartet discrimination score. In our experiments we consider this method and the adaptive method that uses the largest among the six noise-minimizing SR coefficients (*Max-c^{opt}-Quart*).

4.4. Experiments on homogeneous quartets

As with the two-path problem, we tested our methods on quartets simulated under a *homogeneous* model, so that we could compare our methods, which do not assume homogeneity, to an approach that explicitly makes use of the homogeneous assumption to increase its accuracy. Homogeneous quartets have six free parameters: the total rates of the five edges and a shared ti-tv ratio. However, obtaining MLEs of these six parameters together with the quartet split is an extremely computationally intensive task. A more practical and commonly used approach for this problem is based on estimating a shared ti-tv ratio and total rates for the six paths connecting pairs of taxa (Felsenstein, 1989). As in the 3ML method for the two-path problem, estimation is done by examining all six pairwise alignments, and assuming that the six paths are independent (share no edges) but share a common ti-tv ratio. The inferred split is then obtained by applying the FPM to the MLEs of the six total rates. This method, which we term the 7ML method, always produces a valid split, and it uses the homogeneous assumption to reduce the number of free parameters from 12 to 7.

Our experiments used data simulated on a series of homogeneous quartets to compare adaptive distance-based methods, which do not assume homogeneity, to the 7ML method as a point of reference. We considered several archetypical quartet shapes (see Sections 4.5 and 4.6 below). As in Section 3, for each quartet we ran 100,000 simulations of the substitution process along the quartet using sequences of length $n = 500$ bp, and followed the same approach for handling cases with saturated paths. Accuracy rates of different inference methods were recorded using the number of times that the correct split (a, $b|c$, d) is produced. As in our previous simulations, the standard error of our accuracy estimates are bounded from above by 0.16% (see Section 3.2).

4.5. Symmetric quartets

We start by examining symmetric quartets. Symmetric homogeneous quartets have identical external edges, and are thus defined by three parameters: the shared ti-tv ratio, the total rate of the internal edge, and the total rate of all external edges (Fig. 10). In symmetric quartets we have only two types of paths: paths traversing the internal edge (long paths), and paths not traversing the internal edge (short paths). The four long paths have identical substitution parameters (total rate t_{long}) and the two short paths are identical as well (total rate t_{short}). Thus in a way, quartet inference becomes a problem of identifying which entry of the *FPM-sums* vector is associated with the two short paths, making inference in symmetric quartets very similar to inference in a two-path model.

We simulated data for a series of symmetric quartets with ti-tv ratio of $\kappa = 2$, and ${}^{t_{short}}/_{t_{long}} = 0.9$, and compared the inference success rates of the adaptive distance-based methods to the success rates of the 7ML method and the two standard distance-based methods (based on Kimura's original formula and the transversion only formula). As expected, the two adaptive methods, *DiscScore-Quart* and *Max-c^{opt}-Quart*, outperformed the standard distance-based methods, and were as accurate as the 7ML method (Fig. 11). This is interesting, because in the two-path problem, the discrimination score was a considerably better criterion than the maximal coefficient (Fig. 5).

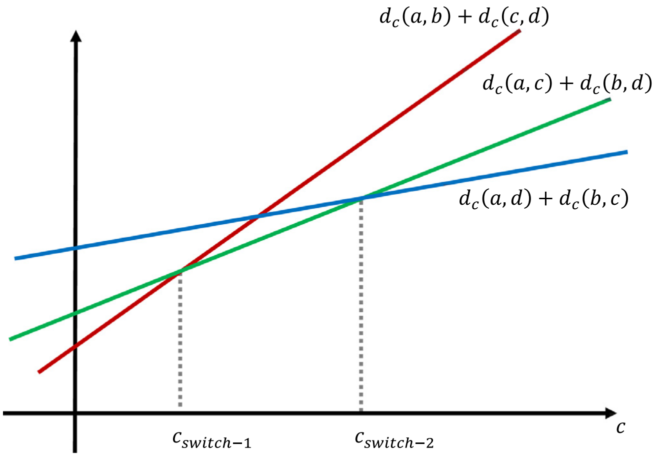

Fig. 9. The figure depicts the three linear functions corresponding to the three entries of *FPMSums*(c; $\hat{\Lambda}$, $\hat{M}$). The inferred split for a given SR function, d_c, is given by the identity of the line which is below the other two in position c. There are at most two switch points between inferred splits. A switch point is an intersection between two of the three linear functions, and there are at most three intersection points between the three lines. Additionally, if there are three intersection points, then one of the lines (green in figure) is below the intersection point of the other two lines (blue and red in figure), implying that this intersection is not a switch point. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

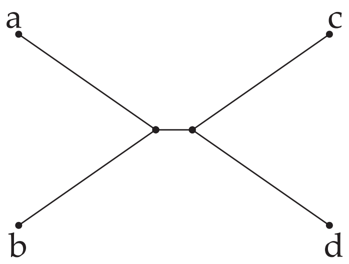

Fig. 10. Symmetric homogeneous quartets are defined by three parameters: the shared ti-tv ratio, the total rate of the internal edge, and the total rate of all external edges. The diameter of the quartet (t_{long}) is defined by the total rate of the longest path in the quartet (between {a, b} and {c, d}).

4.6. Asymmetric quartets

Reconstruction of symmetric quartets is relatively similar to the two-path problem because symmetric quartets have only two distinct types of inter-taxon paths and the two short paths contribute to the same sum in the FPM (Eq. (14)). We thus expect that breaking this symmetry will complicate the inference task. To examine this, we considered quartets in which two of the external edges are longer than the other two. Such asymmetric quartets have three distinct types of edges: the internal edge (typically short), the two long external edges (which are identical), and the two short external edges (also identical). Together with the shared ti-tv ratio, such quartets are defined by four free parameters. The shape of the quartet also depends on whether the identical external edges are on different sides or on the same side of the internal edge. The former type of quartets is associated with the so-called "Felsenstein" zone, which was used to show that the maximum parsimony criterion for phylogenetic inference is statistically inconsistent (Felsenstein, 1978) and also associated with the phenomenon of long-branch attraction (Bergsten, 2005). The latter type is associated with the so-called "Farris" zone, where maximum parsimony was shown to outperform maximum likelihood as a criterion for phylogenetic inference (Siddall, 1998). We thus refer to these

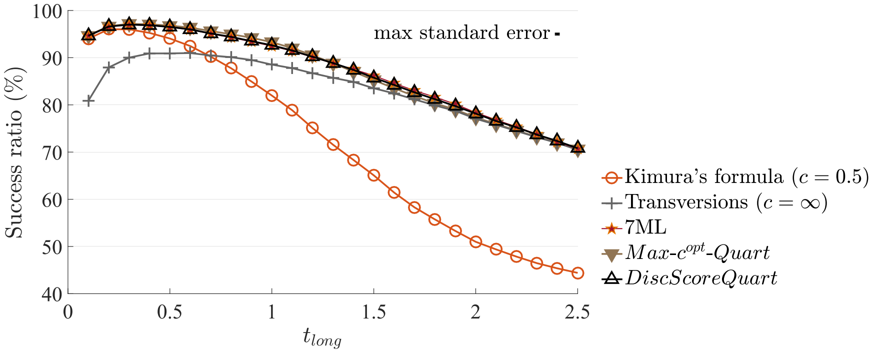

Fig. 11. Experiments on data simulated under symmetric homogeneous K2P quartets with $\kappa = 2$, and $^{t_{short}}/_{t_{long}} = 0.9$. Inference accuracy was measured for the adaptive distance-based methods (*DiscScoreQuart* and *Max-c^{opt}-Quart*) and compared to the accuracy of standard distance-based methods (Kimura's formula and the transversion only formula) and the 7ML method. Both adaptive methods are considerably more accurate than the standard distance-based methods. As in the two-path problem, the method based on the discrimination score (*DiscScoreQuart*) has practically the same accuracy as the 7ML method, but the method based on the largest noise-minimizing SR coefficient (*Max-c^{opt}-Quart*) is relatively better compared to its counterpart in the two-path problem (see Fig. 5). Standard errors for success ratio estimates are less than 0.16% (See Section 4.4).

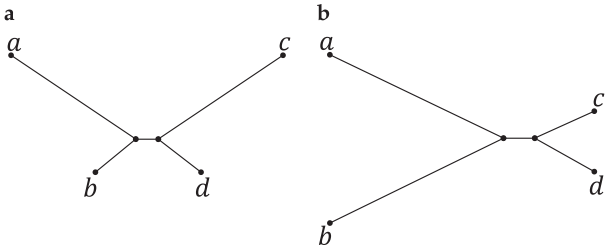

Fig. 12. Two archetypes of asymmetric quartets. In both types, there are two identical long external edges and two identical short external edges. These quartets are thus defined by four free parameters: the shared ti-tv ratio, the total rates of the long external edges, short external edges, and the total rate of the internal edge. In Felsenstein quartets **(a)**, the identical external edges are on different sides of the internal edge, and in Farris quartets **(b)**, the identical external edges are on the same side.

two archetypes of quartets simply as *Felsenstein quartets* (Fig. 12(a)) and *Farris quartets* (Fig. 12(b)).

We conducted experiments on a series of Felsenstein and Farris quartets. We parameterized these quartets according to the total rate of the long path t_{long} containing both long external edges (and the internal edge for Felsenstein quartets). The ratio between the total rate of the internal edge and t_{long} was set to 1:5, and so was the ratio between the total rates of the short and long external edges. As with symmetric quartets, the adaptive distance-based methods are generally better than the standard ones in both types of quartets, but the overall patterns of performance are somewhat more complex (Fig. 13). For instance, both adaptive methods appear to have relatively better performance in Farris quartets than in Felsenstein quartets. This preference appears to be stronger for the *DiscScoreQuart* method than for *Max-c^{opt}-Quart*.

4.7. Topological bias in quartet reconstruction

Our experiments on asymmetric quartets indicate an inherent *topological bias* in quartet reconstruction. Such bias is often attributed to the phenomenon of "long branch attraction", and different phylogenetic inference methods have been shown to have varying degrees of sensitivity to this phenomenon (Siddall, 1998). To examine and quantify this bias more carefully, we conducted a series of additional simulation experiments on quartets with no internal edge. These are quartets that have a star topology with only four edges, all of which are external and connected to a single internal node. Two of the edges are short (identical) and the other two are five times longer (Fig. 14(a)). Note that one of the six inter-taxon paths in the resulting quartet is short (two short edges), one path is long (two long edges), and the other four paths are of intermediate length (short edge and long edge). In the *FPMsums* vector, one entry corresponds to the sum of the short and long paths, and the other two sums correspond to the sums of two intermediate paths. All three sums are identical when considering the true distances, because there is no internal edge. Therefore, given noisy distance estimates, the FPM is expected to return each of the three topologies with equal probability.

We applied different quartet inference methods to these quartets and recorded the number of times that the split (a, $b|c$, d) was returned (the split separating the two long edges from the two short edges). We see that all methods have a bias in favor of this split, and that the bias increases with the scale of the quartet. The two standard distance-based methods (Kimura's formula and the transverion count) have relatively weak bias, remaining below 12% for all scales. As expected, the 7ML method also has very weak bias, only slightly stronger than that observed for the transversion count distance-based method. The adaptive distance-based methods both have somewhat higher degrees of bias, with *Max-c^{opt}-Quart* being somewhat less biased. These results reflect the differences we observed between the performance of the different methods in Felsenstein and Farris quartets: methods with a stronger bias in Fig. 14 show larger differences in their accuracy rates between Fig. 13(a) and (b).

The fact that even standard distance-based methods have some topological bias indicates that this bias is partly due to statistical noise in the distance estimates. This inherent bias makes it more challenging to choose reliable SR functions for inference, because an SR function might appear to confidently infer a quartet split, but this confidence could be a result of biased estimation. Indeed, we see that our adaptive methods are more sensitive to this topological bias.

5. Discussion

Accurate distance-based phylogenetic reconstruction depends on our ability to compare between sums or linear combinations of inter-taxon distances. In this study we examined in detail the basic challenge of comparing the lengths of two independent paths using the framework of adaptive SR functions. The potential of this approach stems from three basic insights: (1) Different SR functions may result in different inference of the identity of the longer path, (2) different SR functions have different levels of distance estimation noise, and (3) we can estimate the expected noise from sequence data. Using these insights, we developed a simple and ef-

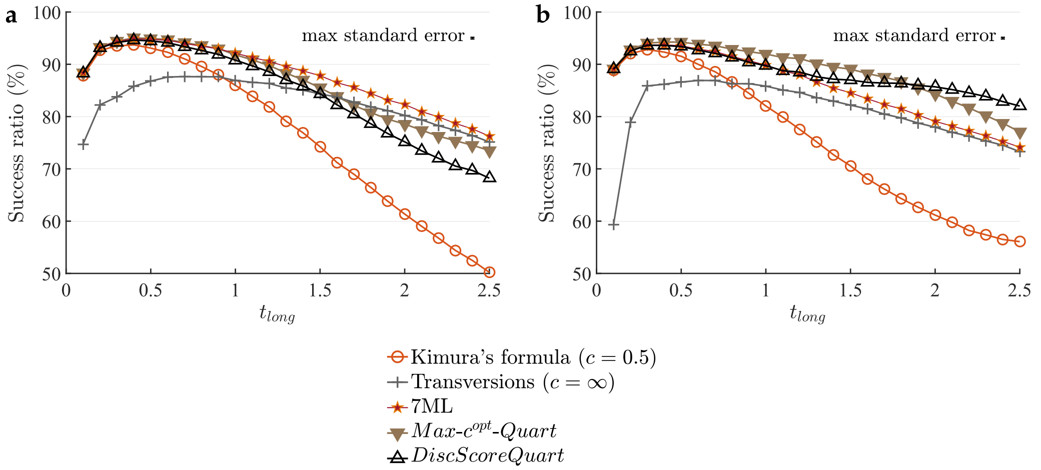

Fig. 13. Experiments on data simulated under asymmetric homogeneous K2P quartets with $\kappa = 2$, and where the ratio between the total rate of the internal edge and the total rate of the long path (t_{long}) was set to 1:5, and so was the ratio between the total rates of the short and long external edges. Data were simulated for Felsenstein quartets (a) and Farris quartets (b). Inference accuracy was measured for the adaptive distance-based methods (*DiscScoreQuart* and *Max-c^{opt}-Quart*) and compared to the accuracy of standard distance-based methods (Kimura's formula and the transversion only formula) and the 7ML method. Differences between performance on symmetric and asymmetric quartets can be attributed to differences in topological bias (see Fig. 14). Standard errors for success ratio estimates are less than 0.16% (See Section 4.4).

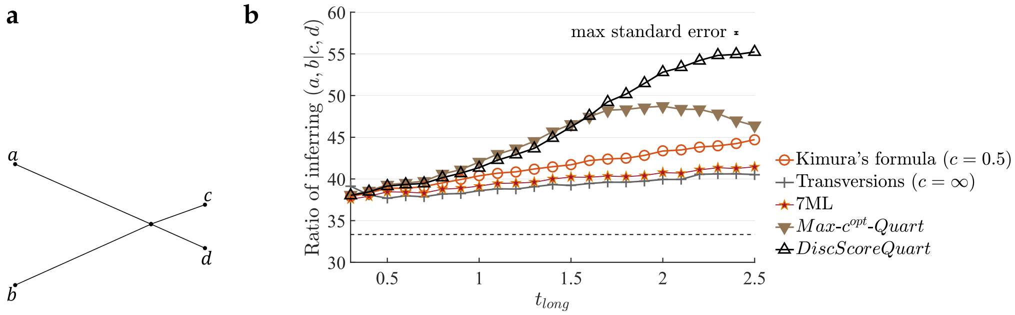

Fig. 14. Experiments on data simulated under asymmetric homogeneous K2P quartets, $\kappa = 2$, with no internal edge, two short external edges and two long external edges (a). (b) Inference rates of the (a, $b|c$, d) split were measured for the adaptive distance-based methods (*DiscScoreQuart* and *Max-c^{opt}-Quart*) and compared to the rates of standard distance-based methods (Kimura's formula and the transversion only formula) and the 7ML method. Because these quartets do not have an internal edge, we expect the inference ratio to be exactly $\frac{1}{3}$. The difference between the inference ratio and 33.3% (horizontal dashed line) measures the bias toward the (a, $b|c$, d) split. All method have some bias, with the adaptive distance-based methods showing the highest levels of bias. Standard errors for inference ratio estimates are less than 0.16% (See Section 4.4).

ficient method to select an SR function for a given comparison. Our proposed method is based on noise-minimizing SR functions introduced by Gronau et al. (2009) and Fisher's linear discriminant. This method is shown to perform as accurately as the ML-based inference for homogeneous models, but unlike ML, our method is not restricted just to homogeneous models and it involves only a few straightforward calculations without the need for numerical optimization. As such, it provides a very appealing alternative to inference based on ML distances, which is the most common practice today.

In the second part of our study, we examined extensions of this basic method to phylogenetic reconstruction, focusing on the fundamental task of quartet inference. Our analysis and experiments indicate some parallels between the two-path problem and the quartet inference problem, but they also highlight some important differences. One basic difference is the fact that paths in a phylogenetic tree are typically overlapping, and thus not independent as assumed in the two-path model. To evaluate the potential influence of dependence, we re-simulated data for our quartet experiments in Figs. 11 and 13 with independent paths (Supplementary Figure S2). These experiments indicate that dependence overall improves the accuracy of all methods. This is because some of the stochastic noise is correlated between overlapping paths, and this correlated noise does not necessarily interfere with *comparison* between path lengths. Importantly, the influence of dependence on our adaptive methods appears to be similar to its influence on the ML-based method, so we conclude that assuming independence does not constitute a significant modeling problem.

Another difference between the two path problem and phylogenetic inference is in the number of values involved in each comparison. The two path problem simply compares two path lengths, but phylogenetic reconstruction algorithms involve steps that typically compare many linear combinations of distances. For instance, the four-point method for quartet resconstruction involves a comparison of three distance-sums. The larger number of compared values could potentially lead to a more complicated structure of ambiguity across different SR functions. However, in Section 4.2 we showed that comparing sums of distances results in a simple convex structure in which the number of switch points is guaranteed to be smaller than the number of compared sums. This result extends beyond sums to general linear combinations of distances, implying that the complexity of comparing linear combinations is strictly bounded by the number of values considered. In practice, we expect that there will be ambiguity only between a few similar values. Thus, the number of compared linear combinations is not expected to constitute a significant bottleneck in the inference task.

The main complication we observed in our quartet inference experiments is that of topological bias in the case of asymmetric quartets. Note that in the two-path problem the tough inference cases involve pairs of paths of similar length. This means that the two estimates being compared have similar statistical noise and we can choose an SR function that reduces noise in both estimates. A similar observation can be made for symmetric quartets. Things get more complicated when we introduce asymmetry. Asymmetric quartets can be tough to infer (short internal edge) even when the underlying paths have very different lengths. This is demonstrated well in the Felsenstein and Farris quartets, which combine short, intermediate and long paths. In these cases, we cannot choose an SR function that will reduced the expected noise for all distances, and having different levels of noise for the three sums in the *FPMsums* vector could contribute to the topological bias shown in Fig. 14.

What appears to make inference difficult in these asymmetric quartets is the fact that we are comparing a pair of balanced paths (of similar length) with a pair of imbalanced paths (one much longer than the other). All methods appear to preferentially "shorten" the unbalanced pair compared to the balanced pair. One of the main challenges of realizing the potential of adaptive distance-based methods in this context is to understand the source of this bias. The self contraction property introduced in Section 3.7 may contribute to bias, since it leads to preferential shortening of certain paths relative to others. Further study of this will lead to development of adaptive methods that better deal with structural bias, and such methods are likely to realize the full potential of the adaptive distance approach by being more statistically robust than standard distance-based methods and much more efficient than ML-based inference.

Acknowledgment

This research was supported by the Israel Science Foundation grant number 509/11.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jtbi.2017.12.022

References

Atteson, K., 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25, 251–278. doi:10.1007/PL00008277.

Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21 (2), 163–193. doi:10.1111/j.1096-0031.2005.00059.x.

Buneman, P., 1971. The recovery of trees from measures of dissimilarity. In: Mathematics in the Archeological and Historical Sciences. Edinburgh University Press, pp. 387–395.

Cavender, J., 1978. Taxonomy with confidence. Math. Biosci. 40, 271–280. doi:10.1016/0025-5564(78)90089-5.

Erdos, P., Steel, M., Szekely, L., Warnow, T., 1999a. A few logs suffice to build (almost) all trees (I). Random Struct. Algor. 14, 153–184. doi:10.1002/(SICI)1098-2418(199903)14:2⟨153::AID-RSA3⟩3.0.CO;2-R.

Erdos, P., Steel, M., Szekely, L., Warnow, T., 1999b. A few logs suffice to build (almost) all trees (II). Theor. Comput. Sci. 221, 77–118. doi:10.1016/S0304-3975(99)00028-6.

Farris, J., 1973. A probability model for inferring evolutionary trees. Syst. Zool. 22, 250–256. doi:10.2307/2412305.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27 (4), 401–410. doi:10.2307/2412923.

Felsenstein, J., 1989. PHYLIP - Phylogeny Inference package (version 3.2). Cladistics 5, 164–166. doi:10.1111/j.1096-0031.1989.tb00562.x.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenic. 7 (2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14 (7), 685–695. doi:10.1093/oxfordjournals.molbev.a025808.

Gronau, I., Moran, S., Snir, S., 2012. Fast and reliable reconstruction of phylogenetic trees with indistinguishable edges. Random Struct. Algor. 40 (3), 350–384. doi:10.1002/rsa.20372.

Gronau, I., Moran, S., Yavneh, I., 2009. Towards optimal distance functions for stochastic substitution models. J. Theor. Biol. 260 (2), 294–307. doi:10.1016/j.jtbi.2009.05.028.

Gronau, I., Moran, S., Yavneh, I., 2010. Adaptive distance measures for resolving K2P quartets: metric separation versus stochastic noise. J. Comput. Biol. 17 (11), 1509–1518. doi:10.1089/cmb.2009.0236.

Hoyle, D.C., Higgs, P.G., 2003. Factors affecting the errors in the estimation of evolutionary distances between sequences. Mol. Biol. Evol. 20 (1), 1–9. doi:10.1093/oxfordjournals.molbev.a004230.

Huson, D.H., Nettles, S.M., Warnow, T.J., 1999. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. J. Comput. Biol. 6 (3–4), 369–386. doi:10.1089/106652799318337.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16 (2), 111–120. doi:10.1007/BF01731581.

Neymann, J., 1971. Molecular Studies of Evolution: A Source of Novel Statistical Problems. In: Gupta, S., Jackel, Y. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, pp. 1–27.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425. doi:10.1093/oxfordjournals.molbev.a040454.

Sattath, S., Tversky, A., 1977. Additive similarity trees. Psychometrika 42 (3), 319–345. doi:10.1007/BF02293654.

Serdoz, S., Egri-Nagy, A., Sumner, J., Holland, B.R., Jarvis, P.D., Tanaka, M.M., Francis, A.R., 2017. Maximum likelihood estimates of pairwise rearrangement distances. J. Theor. Biol. 423, 31–40. doi:10.1016/j.jtbi.2017.04.015.

Siddall, M.E., 1998. Success of parsimony in the four-Taxon case: long-Branch repulsion by likelihood in the farris zone. Cladistics 14 (3), 209–220. doi:10.1006/clad.1998.0063.

Zharkikh, A., 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39 (3), 315–329. doi:10.1007/BF00160155.