Doerr *et al. Algorithms for Molecular Biology* 2012, **7**:22
http://www.almob.org/content/7/1/22

ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Stochastic errors vs. modeling errors in distance based phylogenetic reconstructions

Daniel Doerr[1], Ilan Gronau[2], Shlomo Moran[3*] and Irad Yavneh[3]

Abstract

Background: Distance-based phylogenetic reconstruction methods use evolutionary distances between species in order to reconstruct the phylogenetic tree spanning them. There are many different methods for estimating distances from sequence data. These methods assume different substitution models and have different statistical properties. Since the true substitution model is typically unknown, it is important to consider the effect of model misspecification on the performance of a distance estimation method.

Results: This paper continues the line of research which attempts to adjust to each given set of input sequences a distance function which maximizes the expected topological accuracy of the reconstructed tree. We focus here on the effect of systematic error caused by assuming an inadequate model, but consider also the stochastic error caused by using short sequences. We introduce a theoretical framework for analyzing both sources of error based on the notion of *deviation from additivity*, which quantifies the contribution of model misspecification to the estimation error. We demonstrate this framework by studying the behavior of the Jukes-Cantor distance function when applied to data generated according to Kimura's two-parameter model with a transition-transversion bias. We provide both a theoretical derivation for this case, and a detailed simulation study on quartet trees.

Conclusions: We demonstrate both analytically and experimentally that by deliberately assuming an oversimplified evolutionary model, it is possible to increase the topological accuracy of reconstruction. Our theoretical framework provides new insights into the mechanisms that enables statistically inconsistent reconstruction methods to outperform consistent methods.

Keywords: Phylogenetic reconstructions, Substitution models, Additive substitution rate functions

Introduction

Phylogenetic reconstruction is the task of determining the topology of an evolutionary tree underlying a given set of samples (species) using sequence data extracted from them. This is typically done by assuming some simplified model for DNA sequence evolution, in most cases modeling it as a homogeneous continuous-time Markov process [1-3]. Distance-based reconstruction algorithms tackle this task by first computing a set of $\binom{n}{2}$ pairwise distances between the n input samples and then finding a tree which fits these distances. The distance measures used for this purpose typically reflect the rates of certain substitution events along the evolutionary paths in question. We thus refer to these distance measures as *substitution rate (SR) functions*. The distance-based approach is based on the fact that if the SR function used is *additive* for the underlying substitution model, and the input sequences are sufficiently long, then the topology of the true tree can be efficiently recovered with high probability. However, since the underlying evolutionary model is usually unknown, this assumption is rarely satisfied in practice.

Substitution models used for phylogenetic reconstruction range from the simplest Jukes-Cantor (JC) model [4], through slightly more complex and flexible models, such as Kimura's two-parameter (K2P) model [5] and the Hasegawa-Kishino-Yano model (HKY) [6], to the General Time-Reversible (GTR) model [7,8]. In previous works [9,10] we observed that substitution models which are not too restrictive or too general have many inherently different additive SR functions.

*Correspondence: moran@cs.technion.ac.il
[3]Computer Science Department, Technion - Israel Institute of Technology, Haifa, Israel
Full list of author information is available at the end of the article

We used this basic observation to demonstrate that it is possible to adjust for each given set of DNA sequences a "good" additive SR function, which leads to significantly increased phylogenetic reconstruction accuracy, compared to other additive SR functions. This exploits our ability to predict the *stochastic noise* associated with each SR function. When the SR function used for distance estimation is additive for the underlying substitution model, this stochastic noise is the only cause for inaccurate reconstruction. However, in the scenario, which is very common in practice, where the SR function in use is not additive for the model, an additional systematic bias is introduced in the distance estimates. This systematic bias in distance estimation results in a phylogenetic reconstruction method that might be statistically inconsistent in some cases. In this paper[a], we extend our previous line of research to this scenario, by removing the constraint of additivity. We do this by considering both the stochastic noise and systematic error.

Several previous studies have demonstrated the utility of phylogenetic reconstruction methods that are not generally statistically consistent. The maximum parsimony method has been long known to be inconsistent in some cases [11,12]. However, in other cases it was shown to be more likely to produce accurate reconstructions, compared with the maximum likelihood method [13-15]. More recently, it has been demonstrated that reconstruction accuracy can be improved by deliberately assuming an oversimplified substitution model, when reconstructing a tree using maximum likelihood [16,17]. In the context of distance-based reconstruction, non-additive distance measures have been shown in several cases to lead to improved accuracy when compared with additive measures [18,19]. Overall, these studies provide convincing evidence for the need to consider inconsistent phylogenetic reconstruction methods. However, none of them provide a rigorous framework for characterizing the cases in which inconsistent methods outperform consistent ones.

In this paper we develop a theoretical framework which provides a practical and systematic way to quantify the effect of distance-estimation-bias on the accuracy of distance-based reconstruction. This framework is based on a novel method for measuring the *deviation from additivity* of SR functions. Coupled with the results in [9], this method enables evaluation of both the systematic bias and stochastic noise of SR functions. Such evaluation is important, because there is often a tradeoff between these two sources of error, stemming from the fact that simpler models with fewer parameters (such as JC) have smaller stochastic noise at the expense of greater estimation bias. Our framework allows us to consider this tradeoff when deciding which SR function to use for a given data set. This allows us to characterize a wide range of cases in which an SR function associated with an oversimplified evolutionary model results in increased reconstruction accuracy.

This finding falls in line with previous studies demonstrating the usefulness of phylogenetic reconstruction methods that are not generally consistent. Previous studies have attributed the increased accuracy of inconsistent methods mainly to the fact that these methods have a bias toward reconstructing certain topologies, leading to increased accuracy in cases where the phylogeny being reconstructed has the "favored topology". We notice a similar behavior using our theoretical characterization of non-additive SR functions. However, somewhat surprisingly, we find that non-additive SR functions often have an advantage even when the phylogeny being reconstructed has an "unfavorable topology". This is due to the reduced stochastic noise of the non-additive SR function (compared with it additive alternatives), which compensates for its topological bias.

Our paper is organized as follows. Section "Background" outlines some of the required background and introduces several new concepts that are central in our analysis. Section "Deviation from Additivity in Homogeneous Substitution Models" provides the main analytic results in the paper, and introduces *deviation from additivity* as a measure of distance estimation bias. In that section we prove a general upper bound for this deviation and establish a connection with reconstruction accuracy. We then study deviation from additivity and stochastic error of the JC distance formula when applied to data generated under the K2P model. In Section "Performance of Non affine-additive SR Functions in Quartet Resolution" we study the effect of deviation from additivity and stochastic error on the accuracy of quartet reconstruction. In the case of quartets we can draw a tight connection between the different sources of error in distance estimation and inaccuracy of reconstruction. We present a useful heuristic, based on the so-called Fisher criterion ([20,21]), for comparing the expected accuracy of two SR functions in this context. In Section "Simulations on Hasegawa's Tree" we extend our study to larger trees using experiments on simulated data based on the tree obtained by Hasegawa in [6]. Finally, In Section "Inferring Trees from Genomic Sequences" we demonstrate our approach through a series of experiments reconstructing trees from bacterial gene sequences.

Background

In this section we provide a brief exposition of DNA substitution models and substitution rate functions used for distance estimation. We concentrate on details essential to this study and refer the reader to a previous paper [9] and standard textbooks [1,2] for a more complete survey.

Substitution Models

In this work, a DNA substitution model $\mathcal{M}$ is simply a set of stochastic 4×4 *transition matrices* closed under matrix product (i.e., $\mathbf{P}, \mathbf{Q} \in \mathcal{M} \rightarrow \mathbf{PQ} \in \mathcal{M}$). These matrices serve to describe the substitution process along evolutionary paths in a phylogenetic tree. All substitution models addressed in this paper are time-reversible [7]. A *model tree* in a time reversible substitution model $\mathcal{M}$, or an $\mathcal{M}$*-tree*, is an undirected tree $T = (V, E)$ in which each edge $e \in E$ is associated with a transition matrix $\mathbf{P}_e \in \mathcal{M}$. An $\mathcal{M}$-tree T implies an inter-leaf transition matrix $\mathbf{P}_{ij} \in \mathcal{M}$ for each pair of leaves $\{i, j\} \subset L(T)$, namely $\mathbf{P}_{ij} = \prod_{e \in path_T(i,j)} \mathbf{P}_e$. Most common models are defined using *rate matrices*, which are 4×4 matrices whose off-diagonal elements are non-negative *substitution rates*, and whose rows sum to 0. A stochastic transition matrix $\mathbf{P}$ is obtained from a rate matrix $\mathbf{R}$ through matrix exponentiation: $\mathbf{P} = e^{\mathbf{R}}$.

A common assumption made on the substitution process is that it is *homogeneous* throughout time. This means that all rate matrices in the model are proportional to each other. Such a substitution model is thus termed *homogeneous*, and it is defined by a *unit rate matrix* $\mathbf{R}$ as follows: $\mathcal{M}_{\mathbf{R}} = \{e^{t\mathbf{R}} : t > 0\}$. Note that the definition of the unit rate matrix associated with a given homogeneous model is somewhat arbitrary[b], but once the unit $\mathbf{R}$ is defined, it implies a bijection (or equivalence) between rate matrices in $\mathcal{M}_{\mathbf{R}}$ and the parameter t, which corresponds to evolutionary time. We will make use of this equivalence extensively throughout this paper.

We use the Kimura's two-parameter (K2P) model [5] as a concrete example for demonstrating our approach. A rate matrix in this model is defined by two rate parameters: α, which is the rate of *transition*-type (ti) substitutions (A $\leftrightarrow$ G, C $\leftrightarrow$ T), and β, which is the rate of *transversion*-type (tv) substitutions ({A, G} $\leftrightarrow$ {C, T}). Each K2P rate matrix can be represented as a product of a unit rate matrix, in which $\alpha + 2\beta = 1$, and a scalar t corresponding to *evolutionary time*.

$$\mathcal{M}_{\mathrm{K2P}} = \left\{ e^{t\mathbf{R}_{\alpha,\beta}} \mid t > 0, \alpha \geq \beta > 0, \alpha + 2\beta = 1 \right\} ;$$
$$\mathbf{R}_{\alpha,\beta} = \begin{pmatrix} - & \alpha & \beta & \beta \\ \alpha & - & \beta & \beta \\ \beta & \beta & - & \alpha \\ \beta & \beta & \alpha & - \end{pmatrix} \tag{1}$$

Each unit rate matrix of the K2P model defines a homogeneous sub-model, which is identified by its unique transition-transversion (ti-tv) ratio $R = \frac{\alpha}{2\beta} \geq \frac{1}{2}$. The Jukes-Cantor (JC) model [4] is a special homogeneous sub-model of K2P, in which $R = \frac{1}{2}$ (i.e., $\alpha = \beta$). Although the K2P model is defined in (1) as a union of its homogeneous sub-models, it is important to note that this union is closed under matrix product, implying that K2P adheres to our definition of a proper substitution model. Conversely, some commonly used substitution models, such as GTR and HKY, are defined as a union of homogeneous models, but are not themselves closed under matrix product [22].

Transition matrices in the K2P model have the same symmetric structure as the underlying rate matrices, with two distinct transition parameters: p_α – the probability of a transition-type substitution; p_β – the probability of a transversion-type substitution. The transformations between (α, β, t) and (p_α, p_β) are given by the following equations:

$$\begin{aligned} \alpha t &= -\frac{1}{2}\ln(1 - 2p_\beta - 2p_\alpha) + \frac{1}{4}\ln(1 - 4p_\beta) \\ \beta t &= -\frac{1}{4}\ln(1 - 4p_\beta) \, . \end{aligned} \tag{2}$$

$$\begin{aligned} p_\alpha &= \frac{1}{4}\left(1 + e^{-4\beta t} - 2e^{-2\alpha t - 2\beta t}\right) \\ p_\beta &= \frac{1}{4}\left(1 - e^{-4\beta t}\right) \, . \end{aligned} \tag{3}$$

Substitution rate functions

A *substitution rate (SR) function* for a model $\mathcal{M}$ is a non-negative continuous function $\Delta : \mathcal{M} \rightarrow \mathbb{R}^+$ that maps each transition matrix onto a numerical value of "substitution rate". An SR function Δ induces the following *dissimilarity mapping* over the leaves of an $\mathcal{M}$-tree T: $D^T_\Delta(i, j) = \Delta(\mathbf{P}_{ij})$, for all $\{i, j\} \subset L(T)$. Of particular interest in phylogenetic reconstruction are *additive* SR functions.

Definition 2.1 (Additive SR function)**.** *An SR function Δ is said to be additive for a substitution model $\mathcal{M}$ if for all $\boldsymbol{P}, \boldsymbol{Q} \in \mathcal{M}$, $\Delta(\boldsymbol{PQ}) = \Delta(\boldsymbol{P}) + \Delta(\boldsymbol{Q})$.*

It is often explicitly required that an SR function be additive for the assumed model (see [9]). The evolutionary time, t, typically serves as the standard additive measure in most common substitution models. Throughout this study we follow the special case of K2P, focusing on the two SR functions defined below.

$$\Delta_{\mathrm{K2P}}(p_\alpha, p_\beta) = -\frac{1}{2}\ln(1-2p_\beta-2p_\alpha) - \frac{1}{4}\ln(1-4p_\beta) = \alpha t + 2\beta t \quad = \quad t\,. \tag{4}$$

$$\Delta_{\mathrm{JC}}(p_\alpha, p_\beta) = -\frac{3}{4}\ln\left(1-\frac{4}{3}(p_\alpha+2p_\beta)\right) = -\frac{3}{4}\ln\left(\frac{1}{3}(e^{-4\beta t}+2e^{-2\alpha t-2\beta t})\right). \tag{5}$$

The first SR function, Δ_{K2P}, is the common SR function suggested for the K2P model in [5], and it is clearly additive, as it maps the transition probabilities onto evolutionary time t. The second SR function, Δ_{JC}, maps the transition probabilities onto evolutionary time only in the special case of the JC model where $\alpha = \beta$. Under other homogeneous sub-models of K2P, it is non-additive. This non-additivity is analyzed in details in section Deviation from Additivity in Homogeneous Substitution Models.

Additive metrics, Affine-additive mappings, and Near-additivity

The core idea behind distance-based phylogenetic reconstruction is that a phylogenetic tree T can be accurately and efficiently reconstructed from pairwise distances which are *additive with respect to* T [23,24].

Definition 2.2 (Additive metric)**.** *A metric D defined over the leaf-set $L(T)$ of a tree T is T-additive (or additive w.r.t T), if there exists a positive edge-weighting function $w : E(T) \to \mathbb{R}^+$, such that for each $i,j \in L(T)$, $D(i,j) = \sum_{e \in path_T(i,j)} w(e)$. D is* additive *for a set S if it is T-additive for some tree T where $L(T) = S$.*

It is well known that additive SR functions imply additive metrics: if Δ is an additive SR function for a model $\mathcal{M}$, then for any $\mathcal{M}$-tree T, D^T_Δ (the dissimilarity mapping induced by Δ on T) is a T-additive metric. The inherent difficulty in reconstructing phylogenies using additive SR functions is that computing the implied T-additive metric requires the *exact* values of the inter-taxon transition matrices $\{\mathbf{P}_{ij}\}$, and getting these exact values from alignments of finite length is practically impossible. Therefore, a distance-based reconstruction algorithm is useful in a realistic setting only if it has some robustness to error in distance estimation. In [25], Atteson observed that the topology of a phylogenetic tree T can be accurately (and efficiently) reconstructed from any dissimilarity mapping D which is sufficiently close to a T-additive metric, using certain "robust" distance-based algorithms[c].

Formally, "sufficiently close" is defined by the following relation:

Definition 2.3 (Near-additive mapping)**.** *A dissimilarity mapping D on $L(T)$ is said to be near-additive w.r.t. T iff there exists a T-additive mapping $D^\star$ s.t.*

$$||D, D^\star||_\infty \left(\stackrel{\triangle}{=} \max_{\{i,j\}\subset L(T)} \{|D(i,j) - D^\star(i,j)|\}\right) < \frac{1}{2} w_{\min}(D^\star)\,, \tag{6}$$

where $w_{\min}(D^\star)$ is the minimal weight assigned to an internal edge[d] by the edge weighting function corresponding to the additive metric $D^\star$.

For our results we will be using a generalization of this criterion, in which the mapping $D^\star$ can be any *affine-additive* mapping, defined below.

Definition 2.4 (Affine-additive mapping)**.** *A dissimilarity mapping D' is said to be* affine-additive *w.r.t. a phylogenetic tree T, if there is a T-additive metric D, and scalars $a > 0, b$ s.t. $D' = aD + b$ (i.e., $D'(i,j) = aD(i,j) + b$ for all $\{i,j\} \subset L(t)$).*

As with additive metrics, affine-additive mapping are also associated with edge weights. Let D be a T-additive mapping corresponding to the edge-weighting function $w(\cdot)$. Then the edge weighting function $w'(\cdot)$ corresponding to the affine additive mapping $D' = aD + b$ is given by: $w'(e) = aw(e)$ for all internal edges, and $w'(e) = aw(e) + \frac{1}{2}b$ for all external edges. When b is positive, D' is actually an additive metric, but when b is negative, the weights of external edges implied by $w'(\cdot)$ might be negative, and D' might even yield negative dissimilarities. The generalization of Atteson's theorem to cases where $D^\star$ is affine-additive follows from the observation that the robust distance-based reconstruct algorithms considered by Atteson are invariant to affine transformations of their input distances. From this point on, when we say a dissimilarity mapping D is *near additive*, we mean it satisfies (6) with respect to some affine-additive mapping $D^\star$.

Local consistency

Atteson's result plays a central role in arguing the statistical consistency of distance-based phylogenetic reconstruction. Typically, this is done by assuming that the inter-leaf distances are computed using an SR function Δ which is additive for the underlying substitution model $\mathcal{M}$, as follows:

1. If Δ is additive for $\mathcal{M}$, then for each $\mathcal{M}$-tree T the mapping D^T_Δ defined by $D^T_\Delta(i,j) = \Delta(\mathbf{P}_{ij})$ for all $i,j \in L(T)$, is a T-additive metric.

2. As the length of the input sequences grows, the estimated transition matrices $\{\widehat{\mathbf{P}}_{ij}\}$ converge (w.h.p.) to the true matrices $\{\mathbf{P}_{ij}\}$.
3. When $\{\widehat{\mathbf{P}}_{ij}\}$ are sufficiently close to $\{\mathbf{P}_{ij}\}$, the estimated dissimilarity map $\widehat{D}$ defined by $\widehat{D}(i,j) = \Delta(\widehat{\mathbf{P}}_{ij})$ is sufficiently close to D^T_Δ, and is thus near-additive.
4. The near-additivity of the estimated dissimilarity map $\widehat{D}$ implies accurate topological reconstruction, assuming a robust distance-based algorithm is used.

This line of argument has been used in numerous works studying statistical consistency of distance-based algorithms (e.g., [25-27]), and in all these cases an additive SR function is assumed. Notice, however, that this line of argument remains valid when D^T_Δ is *near additive* w.r.t. T. For instance, consistent reconstruction of any $\mathcal{M}$-tree is guaranteed by using an *affine-additive* SR function Δ', which is an affine transformation of some additive SR function Δ: $\Delta' = a\Delta + b$ (with $a > 0$). An SR function that is not affine-additive in a given substitution model $\mathcal{M}$ does not guarantee consistency across all $\mathcal{M}$-trees, but it still can be consistent for specific $\mathcal{M}$-trees.

Definition 2.5 (Consistent SR function)**.** *An SR function Δ of a substitution model $\mathcal{M}$ is said to be consistent w.r.t. an $\mathcal{M}$-tree T if D^T_Δ is near-additive w.r.t T.*

The main idea endorsed in this paper is that if an SR function only deviates slightly from some SR function which is affine-additive for $\mathcal{M}$, then it might be consistent with respect to many $\mathcal{M}$-trees of interest, and as such should be considered for use in distance based reconstructions.

Deviation from additivity in homogeneous substitution models

In order to assess whether a given SR function Δ is consistent w.r.t. a given model tree T, one has to find an affine-additive mapping $D^\star$ which minimizes the ratio $\frac{||D^T_\Delta, D^\star||_\infty}{w_{min}(D^\star)}$ (see Definition 2.3). This task seems hard in a general setting, but in the special case of homogeneous substitution models it is tractable. Consider a homogeneous substitution model $\mathcal{M}_\mathbf{R}$. The unit rate matrix $\mathbf{R}$ implies a 1-1 mapping between evolutionary time t and rate matrices in $\mathcal{M}_\mathbf{R}$. It is thus useful to view an SR function for $\mathcal{M}_\mathbf{R}$ as a function $\Delta : \mathbb{R}^+ \to \mathbb{R}^+$ which maps the *evolutionary time t* to a dissimilarity measure $\Delta(t)$.

It can be shown that such Δ is affine-additive in the model if and only if $\Delta(t) = at + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$. We define the *deviation* of an SR function Δ from a given affine-additive function $at + b$ in an interval $[\,t_0, t_1]$ as $\frac{1}{a}\max\{|\Delta(t) - at - b| \,:\, t \in [\,t_0, t_1]\,\}$ (the factor $\frac{1}{a}$ normalizes the deviation to units of evolutionary time). The *deviation from additivity* of Δ within $[\,t_0, t_1]$ is defined as the minimum deviation of Δ from any affine-additive function in that interval.

Definition 2.6 (Deviation from additivity)**.** *Let $\Delta : \mathbb{R}^+ \to \mathbb{R}^+$ be an SR function in a homogeneous substitution model. The deviation from additivity of Δ in an interval $[\,t_0, t_1]$ is defined by:*

$$dev(\Delta, [\,t_0, t_1]\,) \stackrel{\triangle}{=} \inf_{a\in\mathbb{R}^+, b\in\mathbb{R}} \left\{ \max_{t\in[t_0,t_1]} \left\{ \frac{|\Delta(t) - at - b|}{a} \right\} \right\}. \tag{7}$$

Lemma 2.7 below presents the basic relation between deviation from additivity and consistency. In Section Performance of Non affine-additive SR Functions in Quartet Resolution we demonstrate the tightness of this relation.

Lemma 2.7. *Let $\mathcal{M}$ be a homogeneous model, and let T be an $\mathcal{M}$-tree with edge lengths (measured in time units) denoted by $\{t_e\}$. Let $t_{\min} = \min\{t_e : e \in T\}$, and assume that all inter-leaf distances in T fall within the interval $[\,t_0, t_1]$. Then any SR function Δ in $\mathcal{M}$ for which $dev(\Delta, [\,t_0, t_1]\,) < \frac{1}{2}t_{\min}$ is consistent w.r.t. T.*

Proof. We need to show that D^T_Δ is near-additive w.r.t. T. Since $dev(\Delta, [\,t_0, t_1]\,) < \frac{1}{2}t_{\min}$, there are $a \in \mathbb{R}^+, b \in \mathbb{R}$ which satisfy

$$\max_{t\in[t_0,t_1]} \left\{ \frac{|\Delta(t) - at - b|}{a} \right\} < \frac{1}{2}t_{\min}.$$

For all $i, j \in L(T)$, denote $t_{ij} = \sum_{e\in path_T(i,j)} t_e$, and let D be the dissimilarity map associated with evolutionary time: $D(i,j) = t_{ij}$. Clearly, D is an additive metric, and the dissimilarity mapping $D' = aD + b$ is an affine-additive mapping. The internal-edge-weights associated with D' are given by $w'(e) = at(e)$ (see discussion following Definition 2.4), implying that $w_{\min}(D') = at_{\min}$. We thus have:

$$\begin{aligned} ||D', D^T_\Delta||_\infty &\le \max_{t\in[t_0,t_1]} \{|\Delta(t) - at - b|\} \\ &< \frac{1}{2}at_{\min} \;=\; \frac{1}{2}w_{\min}(D')\,. \end{aligned}$$

□

An upper bound on the deviation of an SR function Δ from additivity in a given interval $[\,t_0, t_1]$ is implied from the error associated with its linear interpolation $At + B$ within that interval ($A = \frac{\Delta(t_1)-\Delta(t_0)}{t_1-t_0}$ and

$B = \frac{t_1\Delta(t_0)-t_0\Delta(t_1)}{t_1-t_0}$). Figure 1a demonstrates this for Δ_{JC} under a homogeneous sub-model of K2P, and Lemma 2.8 below presents a general upper bound on the deviation from additivity. For this purpose, we assume that the SR function Δ is a monotone increasing continuous function of t with continuous first and second derivatives.

Lemma 2.8. *Let $\Delta : \mathbb{R}^+ \to \mathbb{R}^+$ be an SR function in a homogeneous substitution model, and let $[t_0, t_1]$ be an interval. Let $\Delta_{int}(t) = At + B$ be the linear interpolation of Δ in $[t_0, t_1]$ defined above, and let $F \stackrel{\Delta}{=} \max_{t\in[t_0,t_1]}\{|\Delta''(t)|\}$. Then*

$$dev(\Delta, [t_0, t_1]) \;\leq\; \frac{(t_1 - t_0)^2 F}{16A}. \tag{8}$$

Proof. Let us start by introducing a couple of auxiliary notations:

$$\psi(a, b, t) \;=\; \Delta(t) - at - b$$
$$\psi(a, b) = \max_{t\in[t_0,t_1]} \{|\psi(a, b, t)|\}.$$

We are looking for $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ which minimize $\frac{1}{a}\psi(a, b)$. Let $\psi_{\min} = \min_{t\in[t_0,t_1]}\{\psi(A, B, t)\}$, $\psi_{\max} = \max_{t\in[t_0,t_1]}\{\psi(A, B, t)\}$, and let $b^* = B + \frac{1}{2}(\psi_{\max} + \psi_{\min})$. Then $\psi(A, b^*) = \frac{1}{2}(\psi_{\max} - \psi_{\min})$. A bound for $dev(\Delta, [t_0, t_1])$ will thus follow by showing that $\psi_{\max} - \psi_{\min} \leq \frac{(t_1-t_0)^2F}{8}$.

Since $\Delta_{\mathrm{int}}(t) = At + B$ is a linear interpolation of Δ in $[t_0, t_1]$, we have $\psi(A, B, t_0) = \psi(A, B, t_1) = 0$. Let $t_{\min}$ be an arbitrary point in the interval $[t_0, t_1]$ s.t. $\psi(A, B, t_{\min}) = \psi_{\min} \leq 0$ and let (t_2, t_3) be the maximal open interval in $[t_0, t_1]$ containing $t_{\min}$ in which $\psi(A, B, t) < 0$ (this interval can be empty if $\psi_{\min} = 0$). We define a similar interval (t_4, t_5) in which $\psi(A, B, t) > 0$ around some arbitrary $t_{\max}$ s.t. $\psi(A, B, t_{max}) = \psi_{\max}$. Note that the intervals (t_2, t_3) and (t_4, t_5) are disjoint, and that Δ_{int} is the linear interpolation of Δ in both these intervals (since $\psi(A, B, t_2) = \psi(A, B, t_3) = \psi(A, B, t_4) = \psi(A, B, t_5) = 0$). Therefore, the bound on the error of polynomial interpolation (see, e.g., [28], p. 187) implies that

$$\psi_{\min} \;\geq\; -\frac{(t_3 - t_2)^2 F}{8} \quad \text{and} \quad \psi_{\max} \;\leq\; \frac{(t_5 - t_4)^2 F}{8},$$

Combining these, we get

$$\begin{aligned} dev(\Delta, [t_0, t_1]) &\leq \frac{1}{A}\psi(A, b^*) \;=\; \frac{1}{2A}(\psi_{\max} - \psi_{\min}) \\ &\leq \frac{\left((t_5 - t_4)^2 + (t_3 - t_2)^2\right) F}{16A} \\ &\leq \frac{(t_1 - t_0)^2 F}{16A}. \end{aligned} \tag{9}$$

□

Note. In Appendix 3 we prove that if Δ does not intersect its linear interpolation $\Delta_{\mathrm{int}} = At + B$ within the interval (t_0, t_1), then the function $At+b^*$ mentioned in the proof above is, in fact, the affine-additive function which minimizes the deviation from additivity of Δ in $[t_0, t_1]$. This means that, in such cases, the first inequality in (9) holds in equality. The last inequality in (9) also holds in equality in such cases, because we are guaranteed to have either $[t_2, t_3] = [t_0, t_1]$ (when Δ is bounded from above by its linear interpolation) or $[t_4, t_5] = [t_0, t_1]$ (when Δ is bounded from below by its linear interpolation). Thus, in such a case, the bound of Lemma 2.8 is reduced to the bound on interpolation error (middle inequality in (9)). Cases where Δ does not intersect its linear interpolation are frequent among many SR functions of interest, as this condition holds when Δ is either convex or concave.

Deviation of Δ_{JC} from Additivity in K2P

We now turn to study the deviation of Δ_{JC} from additivity in homogeneous sub-models of K2P with ti-tv ratio $R > \frac{1}{2}$. First, we express Δ_{JC} as a function of the ti-tv ratio R and the time t, using (5) and the relations $\frac{\alpha}{2\beta} = R$ and $\alpha + 2\beta = 1$.

$$\begin{aligned} \Delta_{\mathrm{JC}}(R, t) &= -\frac{3}{4}\ln\left(\frac{1}{3}(e^{-4\beta t} + 2e^{-2\alpha t - 2\beta t})\right) \\ &= -\frac{3}{4}\ln\left(\frac{1}{3}(e^{-\frac{2t}{R+1}} + 2e^{-t\frac{2R+1}{R+1}})\right) \\ &= -\frac{3}{4}\ln\left(\frac{1}{3}e^{-\frac{2t}{R+1}}\left(1 + 2e^{t\frac{2R-1}{R+1}}\right)\right) \\ &= \left(\frac{3}{2(R+1)}\right)t - \frac{3}{4}\ln\left(\frac{1}{3}\left(1+2e^{-t\frac{2R-1}{R+1}}\right)\right). \end{aligned} \tag{10}$$

Note that the homogeneous K2P sub-model with $R = \frac{1}{2}$ is the JC model; in this case the second term of (10) vanishes, leaving $\Delta_{\mathrm{JC}}(\frac{1}{2}, t) = t$. For other homogeneous sub-models of K2P, where $R > \frac{1}{2}$, Δ_{JC} is not affine-additive (i.e., not of the form $at + b$ for $a > 0$), and we can use the result in Lemma 2.8 to bound the deviation of Δ_{JC} from additivity. Denoting $\rho = \frac{2R-1}{R+1}$, we get

$$\frac{\partial\Delta_{\mathrm{JC}}(R, t)}{\partial t} = \frac{3}{2(R + 1)} + \frac{3}{2}\rho\frac{e^{-\rho t}}{1 + 2e^{-\rho t}} > 0. \tag{11}$$

$$\frac{\partial^2\Delta_{\mathrm{JC}}(R, t)}{\partial t^2} = -\frac{3}{2}\rho^2\frac{e^{-\rho t}}{(1 + 2e^{-\rho t})^2} < 0. \tag{12}$$

$$\frac{\partial^3\Delta_{\mathrm{JC}}(R, t)}{\partial t^3} = \frac{3}{2}\rho^3\frac{(1 - 2e^{-\rho t})e^{-\rho t}}{(1 + 2e^{-\rho t})^3}. \tag{13}$$

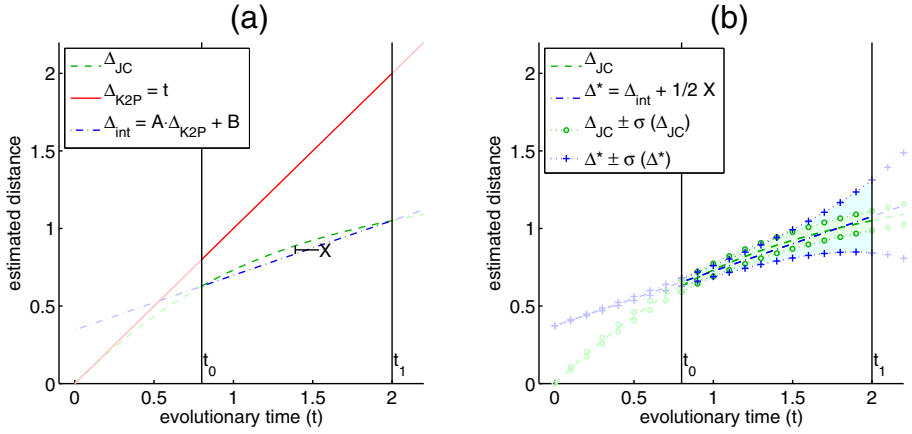

Figure 1 Deviation from additivity and stochastic error. (a) Δ_{JC} is portrayed (green) in the homogeneous sub-model of K2P with $R = 10$ in the interval $t \in [0.8, 2]$. Its linear interpolation in that interval, $\Delta_{int} = At + B$, is plotted in blue, and the maximum difference between the two functions is designated by X. The deviation of Δ_{JC} from additivity in this setting is $\frac{X}{2A}$. **(b)** The affine-additive SR function minimizing its deviation from Δ_{JC} is $\Delta^* = \Delta_{int} + \frac{1}{2}X$. The two SR functions Δ_{JC} and Δ^* are shown with their stochastic error margins, assuming sequence length of 500 bp.

We get that for any given ti-tv ratio $R > \frac{1}{2}$, $\Delta_{JC}(R,t)$ is a concave monotone increasing function, and its second derivative attains a global minimum of $-\frac{3}{16}\rho^2$ at $t = \frac{\ln(2)}{\rho}$. By the note following Lemma 2.8, the deviation of Δ_{JC} from additivity in an interval $[t_0, t_1]$ can be evaluated by computing the linear interpolation $\Delta_{int} = At + B$ of Δ_{JC} in $[t_0, t_1]$, and finding $t \in [t_0, t_1]$ which maximizes $\Delta_{JC}(t) - \Delta_{int}(t)$ (see Figure 1a). A bound on this deviation from additivity can be obtained through Lemma 2.8 by plugging in the slope of the linear interpolation, A, and the maximum value, F, attained by the second derivative of Δ_{JC} in $[t_0, t_1]$. Using Lemma 2.7 and an expression for $dev(\Delta_{JC}(R,t)\ ,\ [t_0, t_1])$, it is possible to map out coherent collections of homogeneous K2P-trees for which Δ_{JC} is guaranteed to be consistent. Each collection is defined by a range of ti-tv ratios $[0.5, R_{max}]$, a range of inter-leaf distances $[t_0, t_1]$, and a lower bound on the weights of internal edges in the tree, given by $t_{min} = 2dev(\Delta_{JC}(R_{max}, t)\ ,\ [t_0, t_1])$.

After determining a collection of trees for which a given non-affine-additive SR function, Δ, is consistent, one can compare the performance of Δ with additive alternatives. In our case, we compare $\Delta = \Delta_{JC}$, which is not affine additive when $R > \frac{1}{2}$, to the standard additive SR function Δ_{K2P}. The potential advantage of Δ_{JC} over Δ_{K2P} lies in its reduced *stochastic noise*. Informally, this occurs because JC relies on the accuracy of estimating a single parameter - the sum $p = p_\alpha + 2p_\beta$, while Δ_{K2P} relies on the accuracy of estimating each of the two parameters p_α and p_β separately. The stochastic noise of an SR function is measured by the *standard deviation* of the statistical estimator associated with it, denoted $\sigma(\Delta_{JC})$ and $\sigma(\Delta_{K2P})$, respectively. We use the result in [9] to get a first order approximation (based on the delta method [29]) of $\sigma(\Delta_{K2P})$ for sequences of length k and model parameters R, t:

$$\sigma(\Delta_{K2P}) \approx \sqrt{\frac{(e^{\frac{4t}{R+1}} - 1) + 4(e^{\frac{2t}{R+1}} - 1) + 2(e^{\frac{4Rt}{R+1}}(e^{\frac{4t}{R+1}} + 1) - 2)}{16k}}. \quad (14)$$

By a similar application of the delta method to Δ_{JC}, we obtain:

$$\sigma(\Delta_{JC}) \approx \sqrt{\frac{p(t,R)\ (1 - p(t,R))}{k\ (1 - \frac{4}{3}p(t,R))^2}}, \quad (15)$$

where k is the sequence length and $p(t,R) = p_\alpha + 2p_\beta = \frac{3}{4} - \frac{1}{4}e^{-\frac{2t}{R+1}} - \frac{1}{2}e^{-\frac{(2R+1)t}{R+1}}$ (see (3)).

Figure 1 provides an illustrative comparison of Δ_{JC} and Δ_{K2P} under the homogeneous sub-model of K2P with ti-tv ratio $R = 10$, and within the inter-leaf time interval of $[0.8, 2]$. Figure 1a shows the deviation of Δ_{JC} from additivity in that setting, using its linear interpolation $\Delta_{int} = At + B$. Note that Lemma 2.8 and the subsequent note imply that $dev(\Delta_{JC}, [0.8, 2]) = \frac{X}{2A}$, where $X = \max_{t \in [0.8,2]}\{\Delta_{JC}(t) - \Delta_{int}(t)\}$. Figure 1b depicts Δ_{JC} in the same setting with its stochastic error margins $(\Delta_{JC} \pm \sigma(\Delta_{JC}))$, alongside its closest affine-additive function $\Delta^* = \Delta_{int} + \frac{1}{2}X$ and its stochastic error margins $(\Delta^* \pm \sigma(\Delta^*))$. These stochastic error margins are determined by assuming a sequence length of 500 bp in the first-order approximations given in (14) and (15), where $\sigma(\Delta^*)$ is given by scaling $\sigma(\Delta_{K2P})$ by the slope A of the linear interpolation. Note how the margins of Δ_{JC} are actually more tightly concentrated around its affine-additive approximation Δ^* than the margins of Δ^*. This implies that,

despite its deviation from additivity in this setting, distances obtained using Δ_{JC} are actually more likely to be near-additive than distances obtained using Δ_{K2P}.

Performance of Non affine-additive SR functions in quartet resolution

The quartet tree is the smallest phylogenetic tree with non-trivial topology. Focusing on quartets enables a close study of the effects of deviation from additivity and stochastic noise on reconstruction accuracy. The topology of a quartet spanning four taxa $\{1,2,3,4\}$ can be represented by the split notation $(ij|kl)$ (where $\{i,j,k,l\} = \{1,2,3,4\}$), indicating that the internal edge of the quartet separates i,j from k,l. All distance based quartet resolution algorithms essentially reduce to the four-point method (FPM) [26,30], which resolves this split using the six observed pairwise distances $\{\widehat{d}_{ij} : \{i,j\} \subset \{1,2,3,4\}\}$: it first partitions the six observed distances into three sums $\widehat{d}_{12}+\widehat{d}_{34}$, $\widehat{d}_{13}+\widehat{d}_{24}$, and $\widehat{d}_{14}+\widehat{d}_{23}$, and then determines the quartet split according to the minimal sum (the sum $\widehat{d}_{ij} + \widehat{d}_{kl}$ corresponds to the split $(ij|kl)$). We will focus on the task of reconstructing homogeneous K2P quartets using FPM with distances $\{\widehat{d}_{ij}\}$ estimated using either Δ_{JC} or Δ_{K2P}. We note that most of our findings easily generalize to more sophisticated homogeneous substitution models, replacing Δ_{JC} by any concave distance function and Δ_{K2P} by some SR function corresponding to the evolutionary time t.

For concreteness, we assume henceforth that the quartet split is $(12|34)$, meaning that the sum of the exact evolutionary times $t_{12} + t_{34}$ is minimal. We start by analyzing the impact of the deviation from additivity of Δ_{JC} on the consistency of quartet resolutions. First, observe that *any* monotone distance function is consistent for quartets in which t_{12} and t_{34} are the smallest interleaf distances - as is the case with symmetric quartets, in which all external edges are of the same length. Therefore, we study two prototypes of asymmetric quartets. The length of the internal edge in both types is t_i, and each type has two long external edges of length t_l, and two short external edges of length t_s.In type A quartets (Figure 2a), the short edges are on one side of the split and the long edges are on the other side. In this case d_{12} and d_{34} are the smallest and largest interleaf distances (resp.). Hence, the concavity of Δ_{JC} increases the separation between the sum $d_{12}+d_{34}$ and the other two competing sums, leading to an expected *improvement* in reconstruction accuracy. The other quartet configuration (type B; Figure 2b) has a short edge and a long edge on both sides of the split. In this case, the interval of interpolation is $[d_{13}, d_{24}]$, and the distance $d_{12} = d_{34}$ is near the center of this interval. Thus the concavity of Δ_{JC} decreases the separation between the sums $d_{13}+d_{24}$ and $d_{12}+d_{34}$ by approximately twice the deviation from additivity of Δ_{JC} in that range.

When the deviation from additivity exceeds half the length of the internal edge, the sum $d_{13}+d_{24}$ becomes the minimal sum, and Δ_{JC} becomes inconsistent. Note that this demonstrates the tightness of the condition stated in Lemma 2.7, and in this sense, type B quartets provide a worst case scenario for quartet resolution by a concave SR function[e].

Next we turn to compare the accuracy of Δ_{JC} with that of Δ_{K2P} when used to reconstruct its "worst case scenario" quartets of type B. Interestingly, Δ_{JC} ends up outperforming Δ_{K2P} on many of these quartets, due to its reduced stochastic noise (as predicted in our discussion revolving around Figure 1b). For example, consider a series of homogeneous K2P quartets of type B with ti-tv ratio $R = 5$, whose edge lengths were set as follows: $t_i = 0.2$, $t_l = 1.0$, and $t_s \in [0.2, 1.0]$. We assessed reconstruction accuracy for both SR functions (Δ_{JC} and Δ_{K2P}) across this series of quartets, by generating 100,000 simulations of the substitution process using 1,000 bp long sequences for each quartet (Figure 3a). Despite its deviation from additivity, Δ_{JC} outperforms the additive SR function Δ_{K2P} on many of these quartets (as long as $t_l/t_s < 3.6$) . Note that as t_s shrinks, the deviation of Δ_{JC} from additivity increases, since the interval $[t_0, t_1]$ expands. This experiment appears to indicate that the deviation of Δ_{JC} from additivity has to be quite large for Δ_{K2P} to outperform it.

Fisher's criterion for separability

We now present a simple and general framework based on the so-called Fisher Criterion (FC) for predicting the relative accuracy of two SR functions in resolving quartets. FC measures the effective separation between normal random variables $X \sim N(\mu_1, \sigma_1)$ and $Y \sim N(\mu_2, \sigma_2)$ using the following measure[f] ([20,21]):

$$FC(X,Y) = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \tag{16}$$

We use FC to measure the separability of the distance sum corresponding to the true split (which should be the minimal sum for consistent SR functions) from the two remaining sums. For the expectation μ of each sum we use the true distances as computed by the SR function on the actual model parameters. For the variance σ^2, we use the sum of the approximate variances of the two distances involved in the sum. We expect that an SR function

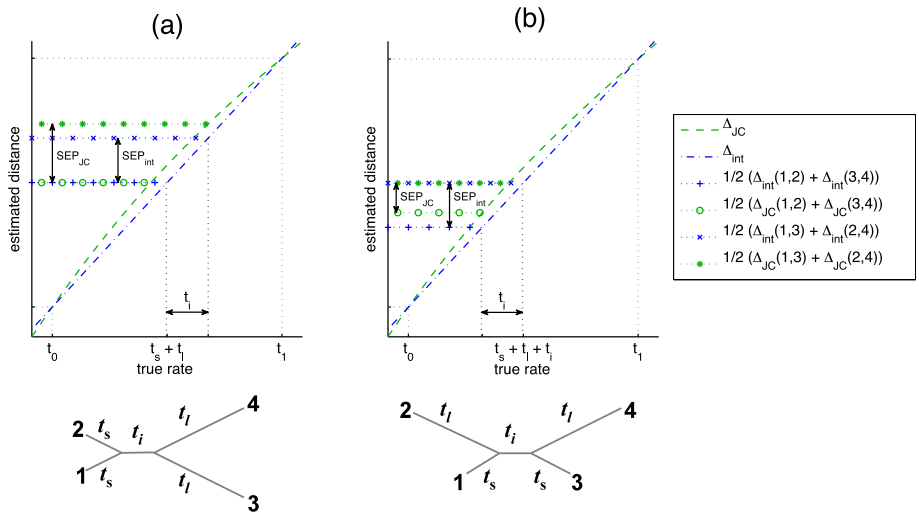

Figure 2 Performance of the Four Point Method using Δ_{JC} on K2P quartets with ti-tv ratio $R = 2$. The concave non affine-additive SR function Δ_{JC} is shown (dashed green line) in the interval $[t_0, t_1]$, where t_0 and t_1 are the smallest and largest of the six pairwise distances (resp.). The dashed blue line shows the linear interpolation $\Delta_{int} = At + B$ of Δ_{JC} in the interval $[t_0, t_1]$. Horizontal dotted lines correspond to half of the two competing sums computed by FPM under the two SR functions (see legend). **(a)** In quartets of type A, $t_0 = t_{12}$ and $t_1 = t_{34}$, and so $\Delta_{int}(1,2) + \Delta_{int}(3,4) = \Delta_{JC}(1,2) + \Delta_{JC}(3,4)$. However, for $i \in \{1,2\}$ and $j \in \{3,4\}$, $\Delta_{int}(i,j) < \Delta_{JC}(i,j)$. Therefore, the deviation from additivity of Δ_{JC} *increases* its FPM separation, denoted SEP_{JC}, compared to the FPM separation SEP_{int} of Δ_{int}. **(b)** In quartets of type B, $t_0 = t_{13}$ and $t_1 = t_{24}$, and so $\Delta_{int}(1,3) + \Delta_{int}(2,4) = \Delta_{JC}(1,3) + \Delta_{JC}(2,4)$. However, $\Delta_{int}(1,2) = \Delta_{int}(3,4) < \Delta_{JC}(1,2) = \Delta_{JC}(3,4)$, and so $\Delta_{int}(1,2) + \Delta_{int}(3,4) < \Delta_{JC}(1,2) + \Delta_{JC}(3,4)$. Therefore, the deviation from additivity of Δ_{JC} *decreases* its FPM separation, denoted SEP_{JC}, compared to the FPM separation SEP_{int} of Δ_{int}. Note that SEP_{int} remains invariant in both types of quartets under fixed t_i whereas SEP_{JC} changes, depending on the type of quartet and the t_s/t_l ratio.

which provides a larger separation of the smallest sum from the two other sums will imply a better reconstruction probability.

We note that FC is not an exact indicator of the separability in our case, because the necessary criteria for this are not satisfied in our model. Namely, the two distance sums are not normally distributed, and they are correlated through the substitution process along the external edges of the quartet. Nevertheless, as Figure 3b suggests, FC turns out to provide a quite reliable comparison of the expected performance of Δ_{JC} and Δ_{K2P} for the quartet series considered in the aforementioned experiment. Figure 3b exhibits for each quartet the FC of Δ_{JC} alongside that of Δ_{K2P}, both associated with the comparison of the true split (12|34) and the "Δ_{JC} favored split" (13|24). As shown, the trends observed in both FC plots closely resemble the trends observed in the reconstruction accuracy plot (Figure 3a), and the the equilibrium point of the FC values of Δ_{JC} and Δ_{K2P} is very close to the equilibrium point of the accuracy of reconstructions of these two functions (near $t_l/t_s = 3.6$).

A useful feature of this framework is the natural way in which it teases apart the stochastic noise from the deviation from additivity. If we denote the numerator of FC by *SEP* (for "separation") and its denominator by *NOISE*, then a comparison of FC estimates between two SR function Δ_1, Δ_2 can be represented as a ratio of ratios:

$$\frac{FC(\Delta_1)}{FC(\Delta_2)} = \frac{SEP(\Delta_1)}{SEP(\Delta_2)} \Big/ \frac{NOISE(\Delta_1)}{NOISE(\Delta_2)}. \quad (17)$$

Figure 4 illustrates how a comparison between the expected performance of Δ_{JC} and that of Δ_{K2P} can be carried out by tracing the *SEP* and *NOISE* ratios along four series of homogeneous K2P quartet: the bottom-left plot corresponds to the quartet series considered in Figure 3; the plot above it corresponds to the same series with ti-tv ratio $R = 2$; the two plots on the right describe two quartet series in which the weight of the short edges is constant $t_s = 0.2$, and the weight of the long edges ranges in $[0.2, 1]$. These four series demonstrate several typical trends in the behavior of the *SEP* and *NOISE* ratios. First, we observe that the *NOISE* ratio decreases (favoring Δ_{JC}) as the diameter of the quartet (t_{24}) increases (it is almost constant in the two series on the left, and monotone decreasing in the series on the right). This is because the diameter provides the major contribution to the stochastic noise (for both Δ_{JC} and Δ_{K2P}), and as it increases, the ratio between the stochastic noise of Δ_{K2P} and Δ_{JC} increases as well. We also observe a natural decrease in the *NOISE* ratio with an increase in the ti-tv ratio (the *NOISE* ratio for $R = 5$ is consistently smaller than for $R = 2$). Concerning

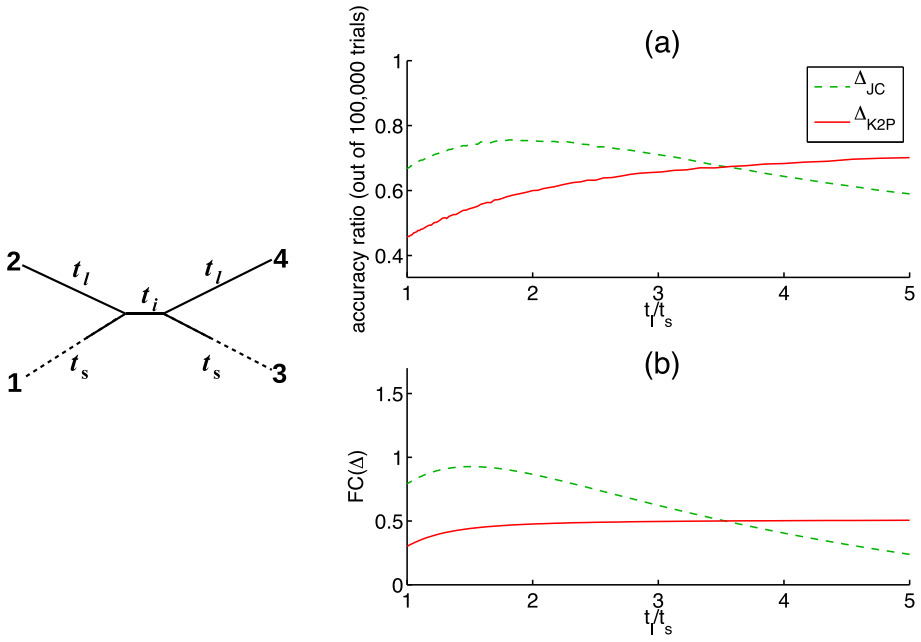

Figure 3 Performance of Δ_{JC} and Δ_{K2P} on a series of quartets of type B. A series of homogeneous K2P quartets is considered (left illustration), with ti-tv ratio of $R = 5$, and edge lengths $t_i = 0.2$, $t_l = 1$, and $t_s \in [0.2, 1]$. **(a)** Reconstruction accuracy using FPM and either Δ_{JC} (dashed green) or Δ_{K2P} (solid red) plotted against t_l/t_s. Accuracy ratio is estimated using 100,000 independent replicates for each value of t_s in the interval $[0.2, 1]$ (in steps of 0.01), with sequence length 1,000 bp. **(b)** Fisher's Criterion (FC) for the sums corresponding to splits (12|34) and (13|24) under either Δ_{JC} (dashed green) or Δ_{K2P} (solid red) plotted against t_l/t_s.

the *SEP* ratio, we see it becomes smaller (favoring Δ_{K2P}) as the quartet becomes more unbalanced (the *SEP* ratio decreases along the X axis in each of the four plots). This is because the deviation of Δ_{JC} from additivity increases as the inter-leaf distance interval $[t_0, t_1] = [t_{13}, t_{24}]$ expands. Deviation of Δ_{JC} from additivity also increases with the ti-tv ratio, as the substitution model further departs from the assumptions of JC (the *SEP* ratio for $R = 5$ is consistently smaller than for $R = 2$).

The two series on the right side of Figure 4 demonstrate well the tradeoff between the effects of stochastic noise and deviation from additivity. In both series, the *SEP* and *NOISE* ratios decrease as the quartets become more unbalanced (due to the trends listed above). However, the rates of decrease of these two ratios are different due to the different ti-tv ratios, and this determines the expected relative performance of the two SR functions across the series. When $R = 2$, the *SEP* ratio decreases at a slower rate than the *NOISE* ratio, and Δ_{JC} is expected to outperform Δ_{K2P} across the entire series. When $R = 5$, the *SEP* ratio decreases at a faster rate than the *NOISE* ratio, and when the quartets are sufficiently unbalanced ($t_l/t_s > 4$) Δ_{K2P} is expected to outperform Δ_{JC}.

Simulations on Hasegawa's Tree

In this section we describe experiments done on simulated data sets generated along the seven-taxon tree assembled by Hasegawa, Kishino, and Yano in 1985 [1,6]. This tree, spanning seven eutherian mammals (Figure 5a), was reconstructed originally using mitochondrial DNA sequences. It has a caterpillar topology (meaning that every internal node is incident to an external edge), and it has long external edges and short internal edges, making it a suitable representative of small phylogenetic trees spanning moderately distant species. These features also make it particularly challenging for distance-based reconstruction.

In our study we used the tree structure and edge lengths to generate simulated data sets. We considered the tree in various scales, by setting the tree diameter (largest inter-taxon path length) to values in the interval $[0.1, 2.0]$. For each scale considered, 10,000 simulations were carried out, where in each simulation 500 bp sequences were evolved along the tree according to a homogeneous K2P substitution model with ti-tv ratio of $R = 2$. For each simulated data set, estimated values of the K2P statistics p_α and p_β, denoted by $\hat{p}_\alpha$ and $\hat{p}_\beta$, were extracted for all $\binom{7}{2}$ pairs of taxa. Subsequently, several distance matrices were computed for each data set by applying different SR functions to these estimated statistics. Reconstruction accuracy was evaluated by applying the Neighbor Joining (NJ) algorithm [31,32] to these distance matrices and recording the *Robinson-Foulds topological distance* (RF) [33] between the reconstructed tree and the Hasegawa tree.

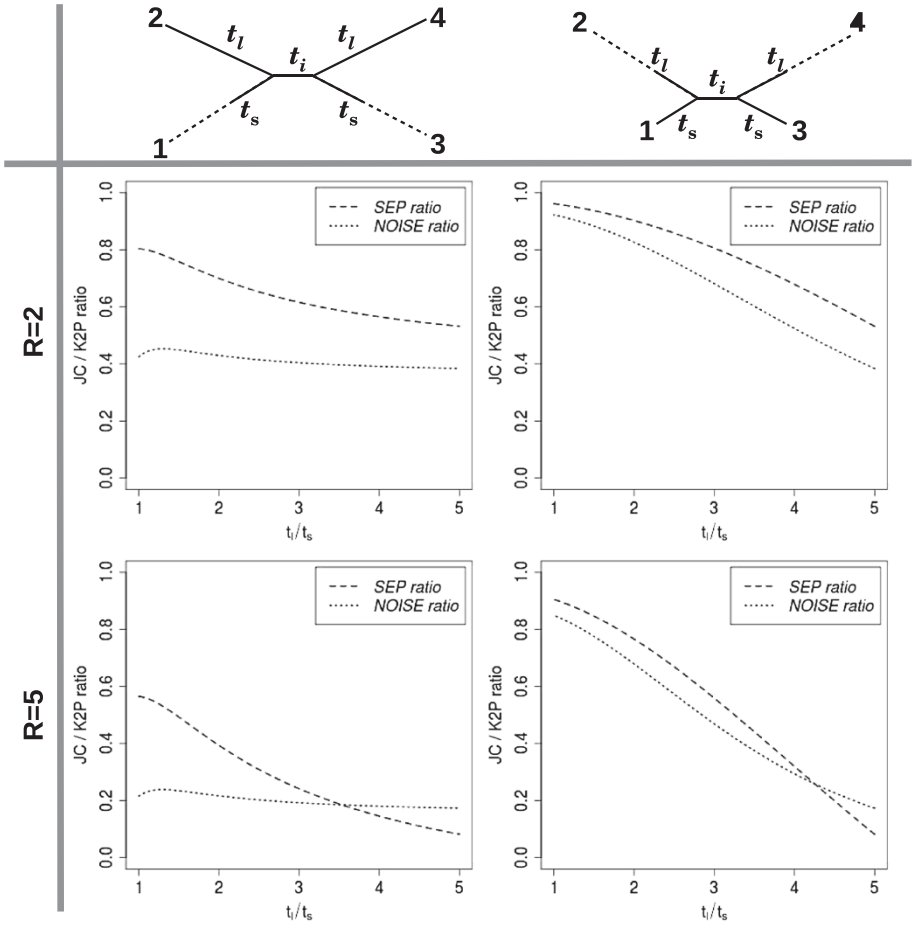

Figure 4 *SEP* and *NOISE* ratios. $SEP(\Delta_{JC})/SEP(\Delta_{K2P})$ (dashed) and $NOISE(\Delta_{JC})/NOISE(\Delta_{K2P})$ (dotted) plotted against t_l/t_s for four series of homogeneous K2P quartets of type B. Top two series have ti-tv ratio of $R = 2$, and bottom two series have ti-tv ratio of $R = 5$. Left two series have external edge lengths $t_l = 1$ and $t_s \in [0.2, 1]$, and right two series have external edge lengths $t_l \in [0.2, 1]$ and $t_s = 0.2$. The length of the internal edge is constant $t_i = 0.2$ in all four series.

Sequence simulation was performed using SeqGen [34] (by choosing the HKY model with uniform base frequencies), and tree reconstruction was performed using the version of NJ implemented in the PHYLIP package [35].

We studied the reconstruction accuracy associated with four different SR functions: Δ_{JC}, Δ_{K2P}, Δ_{tv}, and $\Delta_{R=2}$. The first two are as described in Equations (5) and (4), respectively. The third SR function, Δ_{tv}, considers only tv-type substitutions: $\Delta_{tv}(p_\alpha, p_\beta) = -\frac{1}{4}\log\left(1 - 4p_\beta(t)\right) = \beta t$, and the fourth SR function, $\Delta_{R=2}$, is based on a maximum likelihood (ML) estimator[g] of the time t from the estimated transition probabilities $\hat{p}_\alpha, \hat{p}_\beta$, given that $R = 2$. Informally, this function, which uses knowledge of the true value of R (which is typically unknown to the user), is optimal in our setting, because it has similar stochastic noise as Δ_{JC}, and it is additive since it coincides with Δ_{K2P} when applied to transition probabilities $\hat{p}_\alpha, \hat{p}_\beta$ that are consistent with a ti-tv ratio of $R = 2$.

The performance of these four SR functions is traced across the different tree scales in Figure 5a. For each SR function Δ and scale s, we recorded the average normalized RF distance from the true tree to each of the 10,000 trees reconstructed using Δ. The RF distance was normalized by its maximum value which is twice the number of internal edges in the tree (in our case $2 \times 4 = 8$). As observed previously in [9], Δ_{K2P} performed well in shorter scales, and Δ_{tv} performed well in longer scales. However, both additive SR functions were significantly outperformed in nearly all cases by Δ_{JC}. Surprisingly, Δ_{JC} even slightly outperformed $\Delta_{R=2}$. We speculate that this happened due to a bias similar to the one observed in type A quartets in Section Performance of Non affine-additive SR Functions in Quartet Resolution, improving the performance of concave SR functions such as Δ_{JC} on certain K2P-trees.

To test this hypothesis, we went through a similar experiment with a more symmetric seven-taxon caterpillar tree,

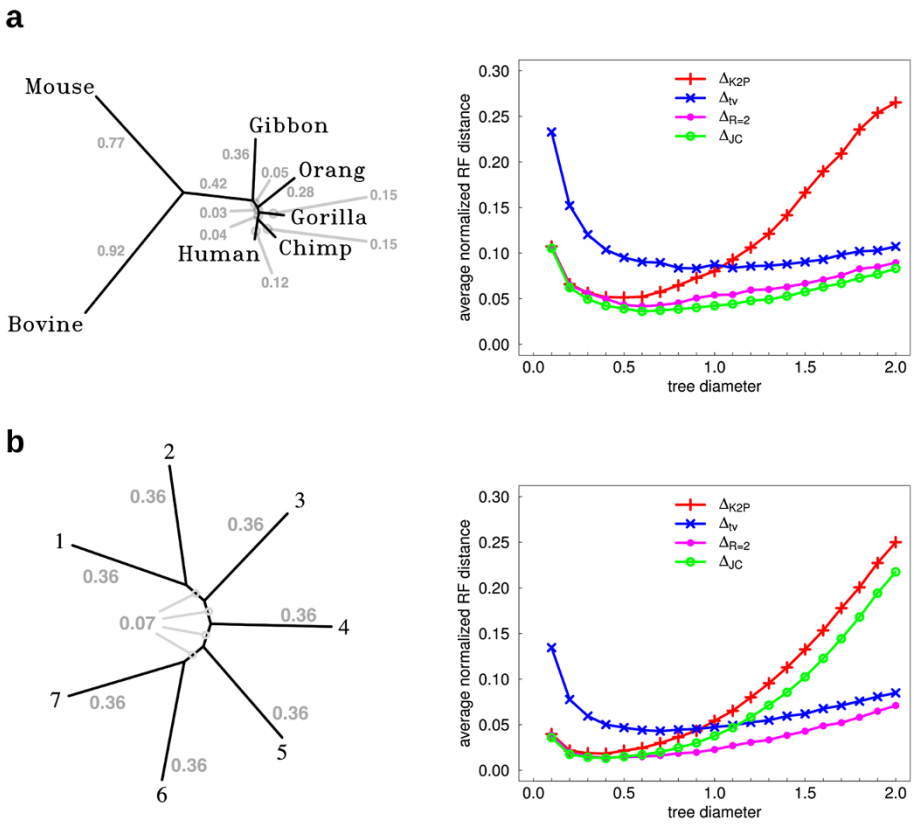

Figure 5 Simulations on Hasegawa's Tree. (a) Reconstruction accuracy of four different SR functions on different scaled versions of Hasegawa's tree [6]. The tree with scaled edge weights is depicted (left) next to the graph (right) plotting reconstruction accuracy of four SR functions. Different scales of the tree are considered, indicated by the diameter of the tree (X axis). Reconstruction accuracy (Y axis) is measured for each scaled tree by the average normalized RF distance between the reconstructed tree and the true tree across 10,000 simulated data sets. Simulations were carried out assuming a ti=tv ratio of $R = 2$ and sequence length of 500 bp. **(b)** A similar plot is shown for a semi-symmetric caterpillar tree.

with internal edges of uniform length t_{int}, and external edges of uniform length $t_{ext} = 5t_{int}$ (Figure 5b). The symmetry of this tree was expected to reduce the effect of the reconstruction bias observed in Hasegawa's tree, and indeed, $\Delta_{\rm JC}$ performed much more poorly on this tree. Despite this fact, $\Delta_{\rm JC}$ still outperformed $\Delta_{\rm K2P}$ in all scales and $\Delta_{\rm tv}$ in the smaller scales ($s < 1.1$).

Inferring trees from genomic sequences

In this section we describe our study comparing various SR functions on genomic DNA sequences. Next to $\Delta_{\rm JC}$ and $\Delta_{\rm K2P}$ we also considered the well known LogDet SR function [36,37], denoted here as $\Delta_{\rm LogDet}$. Extending our study to this setting is challenging in two respects. First of all, unlike the simulated case, the true tree is not known with complete confidence, and accuracy of reconstruction can only be determined by using a well-accepted reference tree that may contain some errors. Secondly, the true substitution model is also unknown and is likely to violate the assumptions of both JC and K2P models and even the relaxed assumptions of the general time-reversible model (in which $\Delta_{\rm LogDet}$ is additive). Hence, we have to assume in this case that $\Delta_{\rm JC}$, $\Delta_{\rm K2P}$, and $\Delta_{\rm LogDet}$ are all non affine-additive, where $\Delta_{\rm JC}$ and $\Delta_{\rm K2P}$ are still likely to exhibit higher deviation from additivity than $\Delta_{\rm LogDet}$, since they make stronger assumptions on the substitution model.

The genomic data set

In building the genomic data set, we made use of a set of 31 *clusters of orthologous groups* (COGs) which was compiled by Ciccarelli et al. and used for inferring phylogenetic relationships amongst a large number of species in [38,39]. These 31 gene families were selected to capture the evolutionary history of the species containing them. This was done in [38] by making sure that the genes in these families have the following properties: (1) they are highly conserved across species, (2) they have a small number of paralogs, and (3) they are weakly affected by horizontal gene transfer. We scanned the NCBI genome database

and found 199 bacterial genomes that contained all annotated COGs. For each of the 31 COGs, we extracted the appropriate protein sequence in each of the 199 bacterial species, choosing an arbitrary paralog in cases of multiple hits. We followed a procedure similar to the one described in [38,39] to obtain reliable multiple-sequence alignments for each COG: we computed a 199-way multiple alignment of the protein sequences of each COG using HMMalign [40] and then mapped each protein sequence back to its coding DNA sequence. The conserved parts of each of the 31 DNA alignments were extracted using GBLOCKS [41] to filter out alignment columns with 50% or more gap symbols. The alignments were manually scanned, and 36 species which contributed a large number of gaps to the alignments were removed from the subsequent analysis. The 31 different alignments were concatenated to form one long 163-way multiple sequence DNA alignment.

For the reference tree we used the phylogenetic tree of microbial species provided by the ARB-SILVA Living Tree Project [42]. This tree, spanning 8,029 species at the time of writing, is based on a widely accepted analysis of the small subunit (SSU) 16S RNA. A subtree spanning our 163 bacterial species was extracted from this tree and treated as the true phylogenetic tree in our analysis.

Reconstruction accuracy for ten-species subsets

We used the base set of 163 species to generate 40,000 random 10-species sub-alignments. The random selection process was guided to generate species subsets corresponding to a wide range of diameter scales (a blind random selection process is biased toward subsets with large diameters). For each of the 40,000 subsets, a 10-way subalignment was extracted from the original 163-way alignment, and in this alignment we extracted only columns corresponding to four-fold degenerate sites that do not have any gap symbol. This is done to make sure the sites used for distance estimation have undergone a substitution process that is as uniform as possible along the different lineages and across the different sites. Each sub-alignment was used to compute three distance matrices – one under Δ_{JC}, one under Δ_{K2P}, and one under Δ_{LogDet}. The latter was calculated by the version that is implemented in the PHYLIP package. The NJ algorithm was then applied to the three matrices and the resulting trees were compared to the true tree (as depicted by the appropriate LTP subtree) according to the RF distance.

As an additional comparison, we used a fourth reconstruction technique. This method (termed BIONJ-GTR) used the BIONJ reconstruction algorithm [43] on distances obtained under the general time-reversible model with invariant sites and Gamma distribution of rates across variant sites (GTR+Γ+I) [8,44].

The PhyML package [45] was used to infer this tree for each of the 40,000 subsets. We selected the GTR+Γ+I model since it was found by the MEGA5 software [46] to provide the best fit to the sequence data. The 40,000 sampled instances were partitioned into eight bins according to the RF distance observed between the BIONJ-GTR tree and the true (LTP) tree, and average RF distances were recorded for each of the three SR functions in each bin. This allowed us to observe trends throughout these 40,000 samples (Figure 6). Of the 40,000 trees inferred under Δ_{JC}, 83.1% showed an equal or lower RF distance than those reconstructed by the BIONJ-GTR method. Moreover, Δ_{JC} outperformed Δ_{K2P} and Δ_{LogDet} on average in all partitions, and Δ_{LogDet} showed by far the worst performance with 48.7% of all reconstructed trees achieving higher RF distances to the reference tree than those inferred by BIONJ-GTR. As with our results on simulated data sets, we see that the SR functions with lower stochastic error but inferior model fit performed best. Unsurprisingly, the GTR+G+I model itself, which was predicted to have the best fit to the sequence data, was often outperformed by the simpler JC and K2P models. Note that the difference in performance between Δ_{JC} and the two other SR functions is greater for subsets that are more accurately reconstructed by the BIONJ-GTR approach (the lower bins). This appears to indicate that oversimplified distance methods are particularly beneficial when the sequence data conveys a stronger phylogenetic signal.

Conclusions

In this paper we explored the basic properties of methods for estimating evolutionary distances, and studied how these properties affect the accuracy of distance-based phylogenetic reconstruction. We considered both the systematic bias and the stochastic noise (variance) of the distance estimators, and examined the tradeoff between these two factors. We focused on the common task of phylogenetic reconstruction under homogeneous substitution models. Assuming homogeneous models simplifies the analytical framework, since in such models each SR function is reduced to a univariate function of the evolutionary time t. However, obtaining accurate estimates of t is still a hard task in this setting, since the unit rate matrix is unknown. An SR function Δ is guaranteed to yield consistent reconstruction across *all* trees in a homogeneous model only if it is additive, meaning that it is a linear function of t. When Δ is not additive, it introduces a systematic bias in distance estimates, which we denoted here as *deviation from additivity*. Some SR functions are only additive in one homogeneous model, whereas others are additive across a wider collection of homogeneous models. This less constrained additivity is typically achieved at a price of increased estimation noise. We studied the tradeoff between "deviation from additivity" and "estimation noise" via a case study where the model tree is a

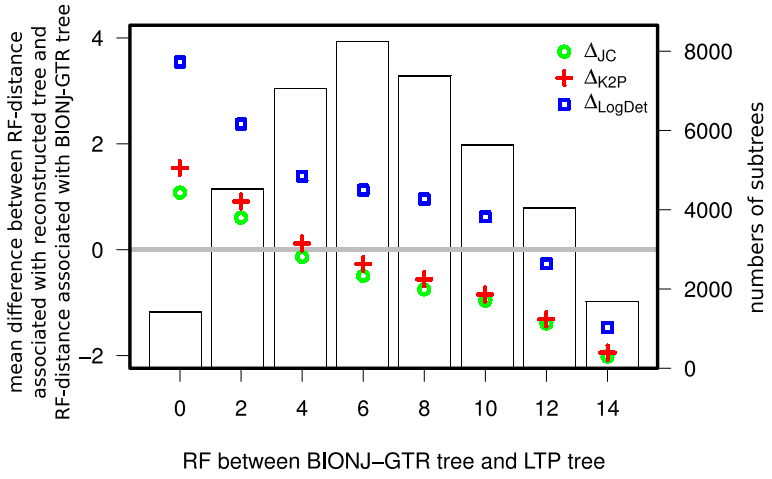

Figure 6 Evaluation against BIONJ-GTR tree. The 40,000 subsets of size 10 were partitioned according to the the RF-distance between the reference LTP tree and the tree reconstructed using BIONJ-GTR (X axis). The (left) Y axis describes the mean difference between the RF-distance associated with a tree reconstructed using a particular SR function (Δ_{K2P}, Δ_{JC}, or Δ_{LogDet}) and the RF-distance associated with the BIONJ-GTR tree. The bar plot in the background depicts the number of subsets in each bin.

homogeneous K2P-tree with an unknown ti-tv ratio R. In this case, Kimura's distance formula Δ_{K2P} is always additive, while the less noisy Jukes Cantor's formula, Δ_{JC}, is additive only when $R = \frac{1}{2}$.

A study of this type requires a way to measure the deviation from additivity of a non-additive SR function Δ in a given range of distances $[t_0, t_1]$. To this end, we introduced the concept of affine-additive distance functions, and defined the deviation from additivity of Δ in $[t_0, t_1]$ as the distance of Δ from its closest affine-additive function in $[t_0, t_1]$. We established a tight connection between this measure and statistical consistency of reconstruction (Lemma 2.7) and derived an upper bound for deviation from additivity in homogeneous models (Lemma 2.8). We applied these results in analyzing the deviation from additivity of Δ_{JC}, and its effect on the accuracy of reconstructing homogeneous K2P-trees. We then showed, both analytically (in Section Deviation from Additivity in Homogeneous Substitution Models) and through experiments on simulated data sets (in Sections Performance of Non affine-additive SR Functions in Quartet Resolution and Simulations on Hasegawa's Tree), that, compared to Δ_{K2P}, it is often better to use the non-additive but less noisy estimates of Δ_{JC}, even when R is quite high. Somewhat surprisingly, we found this to be the case even when the tree being reconstructed has an "unfavorable" topology. Our experiments on bacterial gene sequences (Section Inferring Trees from Genomic Sequences) also indicate that the simple and less noisy SR functions perform better on average than ones that are expected to better fit the true substitution process.

The framework presented in this paper implies a practical way for selecting SR functions which are likely to

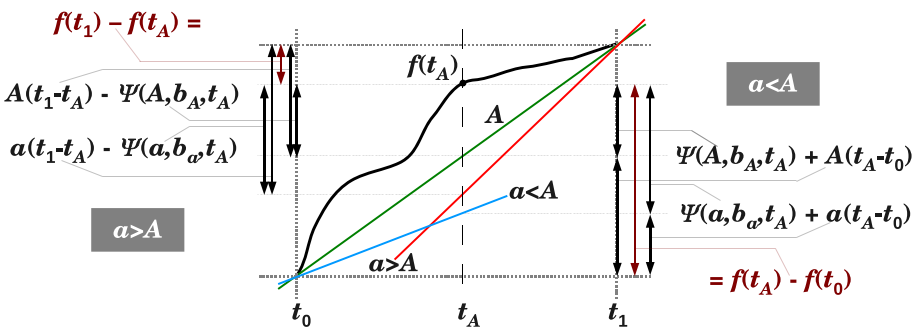

Figure 7 Proof of Lemma 2.9. A function **f(t)** is portrayed (bold) with its linear interpolation $\mathbf{At} + \mathbf{B} = \mathbf{At} + \mathbf{b_A}$ (green) in the interval $[\mathbf{t_0}, \mathbf{t_1}]$, s.t. $\mathbf{f(t)} \geq \mathbf{At} + \mathbf{B}$ for all $\mathbf{t} \in [\mathbf{t_0}, \mathbf{t_1}]$. Equation (18) is illustrated for $a < A$ on the right, and equation (19) is illustrated for $a > A$ on the left.

increase the accuracy of distance estimation. The practicality of the method is drawn from the fact that the criteria by which we select an SR function depend only a relatively crude information about the tree being reconstructed. For instance, in the case of a homogeneous K2P-tree, one can easily obtain from the input sequences rough estimates of both the ti-tv ratio R and the range of inter-leaf times $[t_0, t_1]$. These estimates can then be used to compare the expected accuracies of Δ_{JC} and Δ_{K2P} on the given input, and determine which of them is more likely to yield an accurate phylogeny. For quartets, a tight comparison can be made using the FC-based approach suggested in Section Fisher's Criterion for Separability, and for larger trees, a cruder comparison can be made using a plot like the one presented in Figure 1b. A promising avenue of further research is to extend the FC-based approach to allow tighter prediction of reconstruction accuracy of trees spanning more than four taxa.

Endnotes

[a]This is a WABI 2011 special issue invited paper. Extended abstract of this paper appeared in [47].

[b]Typically, the unit rate matrix is assumed to be the one corresponding to one substitution per site.

[c]Many common distance-based algorithms, such as the Neighbor Joining (NJ) algorithm [31,32], are known to be robust in this sense.

[d]In a tree, edges which touch leaves are *external*, and all other edges are *internal*.

[e]Types A and B quartets represent the *Farris zone* and *Felsenstein zone*, resp. (see, e.g., [1], Chapter 9).

[f]We use here the square root of the criterion commonly used in the literature, because we prefer to think in terms of distances rather than squares of distances. This has no practical influence, since we use FC only for comparing between different choices, not for assessing the quality of a give choice.

[g]This ML estimate is obtained by a simple numerical method for maximizing the likelihood function (see, e.g., [1]).

Appendix

Tightness of Lemma 2.8.

Let $f(t)$ be a (continuous) function on some interval $[t_0, t_1]$. We prove below that if f does not intersect its linear interpolation $At + B$ in that interval, then $dev(f, [t_0, t_1]) = \frac{1}{A}\max_{t\in[t_0,t_1]}\left\{|f(t) - At - b^*|\right\}$. We use the following notations, conforming to the notations in the proof of Lemma 2.8:

$$\psi(a,b,t) = f(t) - at - b$$
$$\psi(a,b) = \max_{t\in[t_0,t_1]}\{|\psi(a,b,t)|\} \quad \psi(a) = \min_{b\in\mathbb{R}}\{\psi(a,b)\}.$$

Lemma 2.9. *Let $f(t)$ be a monotone increasing function in the interval $[t_0, t_1]$ and let $At + B$ be its linear interpolation in $[t_0, t_1]$. If either $f(t) \geq At + B$ for all $t \in [t_0, t_1]$ or $f(t) \leq At + B$ for all $t \in [t_0, t_1]$, then for all $a > 0$, we have $\frac{1}{a}\psi(a) \geq \frac{1}{A}\psi(A)$.*

Proof. We prove the minimality of $\frac{1}{A}\psi(A)$ in the case where $f(t) \geq At + B$ for all $t \in [t_0, t_1]$. The other case (where $f(t) \leq At + B$ for all $t \in [t_0, t_1]$) can be proven in an identical fashion.

For $a > 0$, let b_a be the maximum value of b' s.t. $\psi(a, b', t) \geq 0$ for all $t \in [t_0, t_1]$. Evidently, $\psi(a) = \frac{1}{2}\psi(a, b_a)$. If the linear interpolation of $f(t)$ in $[t_0, t_1]$ is given by $At + B$, then $b_A = B$. We need to show that for every $a > 0$, it holds that $A\psi(a, b_a) > a\psi(A, b_A)$. Let t_A be a point in $[t_0, t_1]$ s.t. $\psi(A, b_A, t_A) = \psi(A, b_A)$. Note that if $a < A$, then the two linear functions $At + b_A$ and $at + b_a$ intersect at $(t_0, f(t_0))$, and if $a > A$, then they intersect at $(t_1, f(t_1))$ (see Figure 7).

For $a < A$, we get the following equality (Figure 7; right):

$$\begin{aligned}\psi(A, b_A, t_A) + A(t_A - t_0) &= f(t_A)\\ -f(t_0) &= \psi(a, b_a, t_A) + a(t_A - t_0)\,.\end{aligned} \tag{18}$$

Hence, since $\psi(a, b_a) \geq \psi(a, b_a, t)$ for every $t \in [t_0, t_1]$, and since $a < A$, we get

$$\begin{aligned}a\psi(A, b_A, t_A) + aA(t_A - t_0) &< A\psi(a, b_a, t_A)\\ + Aa(t_A - t_0) &\Rightarrow\ a\psi(A, b_A) < A\psi(a, b_a).\end{aligned}$$

Similarly, if $a > A$, we get the following equality (Figure 7; left)

$$\begin{aligned}A(t_1 - t_A) - \psi(A, b_A, t_A) &= f(t_1)\\ -f(t_A) &= a(t_1 - t_A) - \psi(a, b_a, t_A)\,,\end{aligned} \tag{19}$$

and $a > A$ implies that

$$\begin{aligned}aA(t_1 - t_A) - a\psi(A, b_A) &> Aa(t_1 - t_A)\\ - a\psi(a, b_a) &\Rightarrow\ a\psi(A, b_A) < A\psi(a, b_a)\,.\end{aligned}$$

□

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors participated in discussing, formulating, and modulating the research. DD performed the simulations and experiments of Sections Simulations on Hasegawa's Tree and Section Inferring Trees from Genomic Sequences. IG and SM initiated and directed the research and drafted the manuscript. IY performed the analysis in Sections Deviation from Additivity in Homogeneous Substitution Models and Section Performance of Non affine-additive SR Functions in Quartet Resolution and contributed to the ideas of the project. All authors contributed to the writing and editing of the manuscript, and all authors read and approved the final manuscript.

Acknowledgements

This research was supported by the Israel Science Foundation (ISF) grant No. 509/11. We also acknowledge the support for the publication fee by the

Deutsche Forschungsgemeinschaft and the Open Access Publication Funds of Bielefeld University. The third author would also like to thanks the Max Planck Institute for Informatics for supporting a visit under which part of this research was carried out.

Author details
[1]Center for Biotechnology, Bielefeld University, Bielefeld, Germany. [2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, USA. [3]Computer Science Department, Technion - Israel Institute of Technology, Haifa, Israel.

Received: 23 December 2011 Accepted: 28 June 2012
Published: 31 August 2012

References

1. Felsenstein J: *Inferring Phylogenies*. Sunderland: MA Sinauer Associated Inc; 2004.
2. Semple C, Steel M: *Phylogenetics*: Oxford University Press ; 2003.
3. Papoulis A, Pillali SU: *Probability, Random Variables and Stochastic Processes*. 4th edition. New York: McGraw Hill Higher Education; 2002.
4. Jukes T, Cantor C: **Evolution of Protein Molecules.** In *Mammalian Protein Metab*. Edited by Munro H. New York: Academic Press; 1969:21–132.
5. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16**(2):111–120.
6. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**(2):160–174.
7. Tavaré S: **Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences.** *Lectures on Mathematics in the Life Sci* 1986, **17:**57–86.
8. Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20:**86–93.
9. Gronau I, Moran S, Yavneh I: **Towards Optimal Distance Functions for Stochastic Substitution Models.** *J Theor Biol* 2009, **260**(2):294–307.
10. Gronau I, Moran S, Yavneh I: **Adaptive Distance Measures for Resolving K2P Quartets: Metric Separation versus Stochastic Noise.** *J Comp Biol* 2010, **17**(11):1391–1400.
11. Felsenstein J: **Cases in which parsimony or compatability methods will be positively misleading.** *Syst Zool* 1978, **27:**401–410.
12. Cavender J: **Taxonomy with confidence.** *Math Biosci* 1978, **40:**271–280.
13. Steel M, Penny D: **Parsimony, likelihood, and the role of models in molecular phylogenetics.** *Mol Biol Evol* 2000, **17:**839–850.
14. Sober E: **A likelihood justification of parsimony.** *Cladistics* 1985, **1:**209–233.
15. Felstenstein J, Sober E: **Parsimony and likelihood: an exchange.** *Syst Zool* 1986, **35:**617–626.
16. Yang Z: **How often do wrong models produce better phylogenies?** *Mol Biol Evol* 1997, **14:**105–108.
17. Bruno WJ, Halpern AL: **Topological bias and inconsistency of maximum likelihood using wrong models.** *Mol Biol Evol* 1999, **16**(4):564–566. [http://www-t10.lanl.gov/billb/BrunoHalpern99.pdf]
18. Zharkikh A: **Estimation of evolutionary distances between nucleotide sequences.** *J Mol Evol* 1994, **39**(3):315–329.
19. Gascuel O, Guindon S: **Efficient Biased Estimation of Evolutionary Distances When Substitution Rates Vary Across Sites.** *Mol Biol Evol* 2002, **19**(4):534–543.
20. Fisher R: **The use of multiple measurements in taxonomic problems.** *Ann of Eugenics* 1936, **7:**177–188.
21. Duda R, Hart P: *Pattern Classification and Scene Analysis*. 1st edition. Hoboken: John Wiley and Sons; 1973.
22. Sumner J, Fernandez-Sanchez J, Jarvis P: **Lie Markov Models.** *J Theor Biol* 2012, **298:**16–31.
23. Buneman P: **The recovery of trees from measures of dissimilarity.** In *Mathematics in the Archeological and Historical Sciences*. Edited by Hodson F, Kendall D, Tautu P: Edinburgh University Press; 1971:387–395.
24. Sattath S, Tversky A: **Additive similarity trees.** *Psychometrica* 1977, **42**(3):319–345.
25. Atteson K: **The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction.** *Algorithmica* 1999, **25:**251–278.
26. Erdos P, Steel M, Szekely L, Warnow T: **A few logs suffice to build (almost) all trees (I).** *Random Struct Algorithms* 1999, **14:**153–184.
27. Erdos P, Steel M, Szekely L, Warnow T: **A few logs suffice to build (almost) all trees (II).** *Theoret Comput Sci* 1999, **221:**77–118.
28. Johnson L, Riess R: *Numerical Analysis*. Boston: Addison Wesley; 1977.
29. Oehlert G: **A note on the delta method.** *Am Statistician* 1992, **46:**27–29.
30. Zaretskii K: **Constructing a tree on the basis of a set of distances between the hanging vertices.** *Uspekhi Mat Nauk* 1965, **20**(6):90–92. [In Russian].
31. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406–425.
32. Studier J, Keppler K: **A note on the neighbor-joining algorithm of Saitou and Nei.** *Mol Biol Evol* 1988, **5**(6):729–731.
33. Robinson F, Foulds R: **Comparison of phylogenetic trees.** *Math Biosci* 1981, **53:**131–147.
34. Rambaut A, Grass NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Comput Appl Biosci* 1997, **13**(3):235–238.
35. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164–166.
36. Steel M: **Recovering a tree from the leaf colourations it generates under a Markov model.** *Appl Math Lett* 1994, **7**(2):19–24.
37. Lockhart P, Steel M, Hendy M, Penny D: **Recovering evolutionary trees under a more realistic model of sequence evolution.** *Mol Biol Evol* 1994, **11**(4):605–612.
38. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward Automatic Reconstruction of a Highly Resolved Tree of Life.** *Science* 2006, **311**(5765):1283–1287.
39. von Mering, C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P: **Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments.** *Science* 2007, **315**(5815):1126–1130.
40. Durbin R, Eddy SR, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*: Cambridge University Press; 1999.
41. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Bio* 2007, **56:**564–577.
42. Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R: **Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses.** *Syst Appl Microbiol* 2010, **33:**291–299.
43. Gascuel O: **BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.** *Mol Biol Evol* 1997, **14**(7):685–695.
44. Rodriguez F, Oliver JL, Marin A, Medina JR: **The general stochastic model of nucleotide substitution.** *J Theor Biol* 1990, **142:**485–501.
45. Guindon S, Gascuel O: **A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696–704.
46. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
47. Doerr D, Gronau I, Moran S, Yavneh I: **Stochastic Errors vs. Modeling Errors in Distance Based Phylogenetic Reconstructions.** In *Algorithms in Bioinformatics, Volume 6833 of Lecture Notes in Computer Science*. Edited by Przytycka T, Sagot MF. Berlin / Heidelberg: Springer; 2011:49–60.

doi:10.1186/1748-7188-7-22
Cite this article as: Doerr *et al.*: Stochastic errors vs. modeling errors in distance based phylogenetic reconstructions. *Algorithms for Molecular Biology* 2012 **7**:22.