Stochastic Errors vs. Modeling Errors in Distance Based Phylogenetic Reconstructions

Daniel Doerr Ilan Gronau Shlomo Moran Irad Yavneh

July 6, 2011

Abstract

Distance based phylogenetic reconstruction methods use the "evolutionary distances" between species in order to reconstruct the tree spanning them. 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 accuracy of the reconstructed tree. We demonstrate both analytically and experimentally that by deliberately assuming an oversimplified evolutionary model, it is possible to increase the accuracy of reconstruction.

1 Introduction.

Distance based reconstructions of phylogenetic trees from a set of n genetic sequences (DNA or protein) usually consist of the following four steps: (1) a substitution model of sequence evolution is assumed; (2) a *substitution rate* (SR) function Δ is selected (Δ typically corresponds to *additive* distances in the assumed model); (3) the $\binom{n}{2}$ interspecies distances defined by Δ are estimated from alignments of the input sequences; (4) a tree spanning the n species which best fits the estimated distances is constructed.

Models of DNA evolution used for step (1) are usually based on Markovian processes [22]. Among the more common models are the Jukes-Cantor (JC) model [17], Kimura's two-parameter (K2P) model [18], the Tamura-Nei model [34], the Hasegawa-Kishino-Yano (HKY) model [15] and the General Time-Reversible (GTR) model [36, 19]. These models differ in the degree of symmetry imposed on the associated rate matrices $\mathbf{R}$, with the GTR model imposing the least symmetry. Methods and software for selecting the most likely model for a given set of aligned sequences can be found in [35, 14]. Further information on substitution models can be found in [6, 8, 28].

There are two main sources for inaccuracies in the 4-steps phylogentic reconstruction described above: (a) a wrong model chosen in (1) could imply that the function Δ selected in (2) is not additive for the true model; (b) *stochastic errors* associated with the estimation of distances from alignments of finite length in (3). In previous works [12, 13] we have shown that most common DNA substitution models (eg, all the above mentioned models except for Jukes-Cantor) have many different additive SR functions with different patterns of stochastic errors. We demonstrated that selecting a function that is expected to be least noisy for the given input leads to significant improvement in the accuracy of the reconstructed tree. In this paper we extend this line of research to cases where the selected model is not the true model. Somewhat surprisingly, we show both analytically and via experiments on real and simulated data, that by deliberately assuming an oversimplified evolutionary model and using a non-additive but less noisy SR function, it is possible to increase the accuracy of reconstruction. In a sense, this is the "distance-methods" analogue of the following well known phenomenon: the Maximum Parsimony reconstruction method, which is not statistically consistent in general, provides a higher reconstruction accuracy in certain cases compared to reconstruction methods that are statistically consistent (see, e.g., [31, 29, 9]).

In Section 2 we present the required background, and introduce *affine-additive* mappings - a simple generalization of additive distances needed for our analysis. In Section 3 we present the concept of *deviation from additivity* which measures the deviation of an SR function which is not affine-additive from the closest affine-additive one, and then prove a general upper bound on this deviation. Then we compare this deviation with the stochastic noise in the case that the true model is Kimura's two-parameter [18] and the simplified model is Jukes-Cantor. Section 4 demonstrates the possible advantage of using oversimplified model on the reconstruction of quartets by the four-points method ([39, 5]), and then presents a useful heuristic, based on Fisher's linear discriminant ([10, 2]), for identifying scenarios in which such oversimplification is useful. In Section 5 we extend this study to reconstruction of phylogenies based on the Hasegawa tree [15, 8], and Section 6 demonstrates our approach in reconstruction of phylogenies from real biological sequences.

2 Background.

2.1 Substitution Models.

We start with a brief presentation of the concepts used in this paper. A more detailed exposition of these concepts can be found in [12] and in standard textbooks [8, 28].

rewrote this subsection SM110629

A DNA substitution model $\mathcal{M}$ consists of a set of stochastic 4×4 *transition matrices* (describing possible substitution patterns) 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. Each transition matrix has a unique (row) stationary vector Π_{stat} such that $\Pi_{\text{stat}}\mathbf{P} = \Pi_{\text{stat}}$. $\mathbf{P}$ is *time-reversible* if $\mathbf{\Pi}_{\text{stat}}\mathbf{P}$ is a symmetric matrix, where $\mathbf{\Pi}_{\text{stat}}$ be the diagonal matrix representation of Π_{stat}. Most common substitution models assume that transition matrices are time-reversible.

In this paper we consider time-reversible substitution models which are based on Markovian processes [22], and in which all transition matrices share the same stationary vector Π_{stat}. In such models there is a natural notion of *evolutionary time* (or just *time*), based on the notion of unit rate matrix, defined as follows. A *rate matrix* is a 4×4 matrix whose off-diagonal elements are non-negative *substitution rates*, and whose rows sum to 0. A rate matrix $\mathbf{R}$ is a *unit rate matrix* iff $trace(\mathbf{R\Pi}_{\text{stat}}) = -1$. Each transition matrix in the model is given as a matrix exponentiation $\mathbf{P}(\mathbf{R}, t) = e^{t\mathbf{R}}$ where $\mathbf{R}$ is a unit rate matrix and t is (evolutionary) time. A *homogeneous* substitution model is a model defined by a *fixed* unit rate matrix $\mathbf{R}$, that is: $\mathcal{M}_{\mathbf{R}} = \{e^{t\mathbf{R}} \ : \ t \in \mathbb{R}^+\}$. Specifically, in a homogenous model there is one to one correspondence between transition matrices and evolutionary time. Homogeneous models are useful in practice, and in fact are assumed by common phylogenetic software packages like PHYLIP [7].

A *model tree* in a substitution model $\mathcal{M}$ is an undirected tree $T = (V, E)$ in which each edge $e \in E$ is associated with a transition matrix $\mathbf{P}_e(t_e) \in \mathcal{M}$. A model tree T implies an inter-leaf transition matrix $\mathbf{P}_{ij} \in \mathcal{M}$ for each pair of leaves $\{i, j\} \subset L(T)$.

The Kimura's two-parameter (K2P) model [18] is defined by rate matrices $\mathbf{R}$ with two parameters: α, which is the rate of *transition*-type (ti) substitutions ($\mathtt{A} \leftrightarrow \mathtt{G}$, $\mathtt{C} \leftrightarrow \mathtt{T}$), and β, which is the rate of *transversion*-type (tv) substitutions ($\{\mathtt{A},\mathtt{G}\} \leftrightarrow \{\mathtt{C},\mathtt{T}\}$). The unique stationary vector in this model is uniform: $\Pi_{\text{stat}} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Hence $\mathbf{R} = \mathbf{R}_{\alpha,\beta}$ is a unit rate matrix iff $\alpha + 2\beta = 1$.

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

Transition matrices in the K2P model are defined by two parameters: $p_\alpha(t)$ indicating the probability of a transition-type substitution and $p_\beta(t)$ indicating the probability of a transversion-

type substitution in time t. The transformations between $(\alpha t, \beta t)$ and $(p_\alpha(t), p_\beta(t))$ are given by:

$$\alpha t = -\frac{1}{2}\ln(1-2p_\beta(t)-2p_\alpha(t)) + \frac{1}{4}\ln(1-4p_\beta(t)) \qquad \beta t = -\frac{1}{4}\ln(1-4p_\beta(t)) \ . \tag{2}$$

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

A homogeneous sub-model of $\mathcal{M}_{\mathrm{K2P}}$ is defined by the set of all K2P rate matrices that share the same ti-tv ratio $R = \frac{\alpha}{2\beta} \geq \frac{1}{2}$. The Jukes-Cantor (JC) model [17], $\mathcal{M}_{\mathrm{JC}}$, is a special homogeneous sub-model of $\mathcal{M}_{\mathrm{K2P}}$ in which $\alpha = \beta$ (and consequently $p_\alpha(t) = p_\beta(t)$), corresponding to a ti-tv ratio of $R = \frac{1}{2}$.

2.2 Substitution Rate Functions.

A *substitution rate (SR) function* for a model $\mathcal{M}$ is a non-negative continuous[1] function $\Delta : \mathcal{M} \to \mathbb{R}^+$ 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 a model tree T in $\mathcal{M}$: $D^T_\Delta(i,j) = \Delta(\mathbf{P}_{ij})$, 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 $\mathbf{P}, \mathbf{Q} \in \mathcal{M}$, $\Delta(\mathbf{PQ}) = \Delta(\mathbf{P}) + \Delta(\mathbf{Q})$.*

In this paper we study the possible advantages of using non-additive SR functions for phylogenetic reconstruction. In the case study of $\mathcal{M}_{\mathrm{K2P}}$, we will analyze two specific SR functions:

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

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

Δ_{K2P} is the common (additive) SR function used in the general context of $\mathcal{M}_{\mathrm{K2P}}$, as suggested in [18], because it estimates the evolutionary time t. Δ_{JC} coincides with Δ_{K2P} when the ti-tv ratio is $R = \frac{1}{2}$, but it is non-additive in all other homogeneous sub-models of $\mathcal{M}_{\mathrm{K2P}}$.

2.3 Consistent Reconstruction and Near Additivity.

The core idea behind distance-based phylogenetic reconstruction is that a phylogenetic tree T can be accurately and efficiently reconstructed given pairwise distances which are *additive with respect to T* [27].

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 an edge-weighting function $w : E(T) \to \mathbb{R}^+$ which assigns strictly positive weights to all edges, 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 model tree $T \in \mathcal{M}$, D^T_Δ (the dissimilarity mapping induced by Δ on T) is a T-additive metric.

The inherent difficulty in reconstructing phylogenies from additive SR functions is that computing the implied T-additive metric requires the *exact* values of the inter-taxon transition matrices $\{\mathbf{P}_{ij}\}$,

[1] Continuity is assumed under any common matrix norm.

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 [1], Atteson observes that the topology of a phylogenetic tree T can be accurately (and efficiently) reconstructed from any dissimilarity mapping D which satisfies the following: For some T-additive metric $D^\star$, the maximal difference between D and $D^\star$, i.e. $\max_{i,j}\{|D(i,j) - D^\star(i,j)|\}$, is smaller than half the length of the shortest *internal* edge[2] in T. For our results we need a simple generalization of this criterion, in which $D^\star$ can be any *affine-additive* mapping, defined below.

changed description of Atteson's SM110629

Definition 2.3 (Affine-additive mapping)**.** *Let D be a T-additive metric over a leaf-set $L(T)$, associated with edge weights $\{w(e) : e \in T\}$. For each $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$, the* affine *T-additive mapping D_{ab} is given by $D_{ab}(i,j) = aD(i,j) + b$. A mapping $D^\star$ is affine-additive if it is affine T-additive for some T.*
An SR function $\Delta^\star$ is affine-additive *if, for some $a \in \mathbb{R}^+, b \in \mathbb{R}$, it holds that $\Delta^\star(M) = a\Delta(M) + b$ for some additive SR function Δ (and for all model matrices M).*

Note that if $\Delta^\star$ is an affine-additive SR function, then $D^T_{\Delta^\star}$ is an affine T-additive mapping for each model tree T.

Let D_{ab} be any affine-additive mapping, and let w_{ab} be the edge weighting defined by $w_{ab}(e) = aw(e)$ for internal edges, and $w_{ab}(e) = aw(e) + \frac{1}{2}b$ for external edges. Let further $D_{ab}(i,j) = \sum_{e \in path_T(i,j)} w_{ab}(e)$. Note that if $w_{ab}(e) > 0$ for all $e \in T$, then D_{ab} is in fact an additive metric. However in general it is possible that $w_{ab}(e) < 0$ for some external edges e, and hence it is possible that $D_{ab}(i,j) < 0$ for some $i,j \in L(T)$. Nevertheless, the internal edge weights defined by D_{ab} are positive and proportional to these defined by D, and therefore Atteson's criterion applies to D iff it applies to D_{ab}. This implies the following extension of the concept of "near-additive metric" of [1]:

Definition 2.4 (Near-additive mapping)**.** *A dissimilarity mapping D on $L(T)$ is said to be* near-additive *w.r.t. T iff there exists an affine T-additive mapping $D^\star$ s.t.*

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

where $w_{\min}(D^\star)$ is the minimal weight assigned to an internal edge by the edge weighting function corresponding to the affine-additive mapping $D^\star$.

changed the discussion here SM

Assume now that Δ is an SR function s.t. D^T_Δ is a near-additive metric for a model tree T. Then Δ implies a *statistical consistent* distance-based reconstruction of T by the following line of argument:

1. As the input sequences length grows, the estimated transition matrices $\widehat{\mathbf{P}}_{ij}$ converge (w.h.p.) to the exact matrices $\mathbf{P}_{ij}$.
2. Hence, the estimated values $\Delta(\widehat{\mathbf{P}}_{ij})$ converge (w.h.p) to the exact values $\Delta(\mathbf{P}_{ij})$.
3. The near-additivity of the mapping $\{\Delta(\mathbf{P}_{ij}) : i,j \in L(T)\}$ implies that, for long enough sequences, the mapping $\{\Delta(\widehat{\mathbf{P}}_{ij}) : i,j \in L(T)\}$ are also near-additive.
4. Finally, the near-additivity of $\{\Delta(\widehat{\mathbf{P}}_{ij})\}$ guarantees accurate reconstruction of the correct tree topology.

By the above, near-additivity implies statistical consistency. This suggests the following:

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

[2] An edge e is called *internal* if both its endpoints are internal vertices, otherwise it is called *external*.

Clearly, an SR function Δ which is affine-additive for $\mathcal{M}$ is consistent for all model trees in $\mathcal{M}$. However, Δ might be consistent for many model trees of interest even if it is not affine-additive. In fact, as we demonstrate later, a deliberate selection of consistent functions which are not affine-additive but have smaller stochastic noise (since they assume an oversimplified model) often increases the accuracy of the reconstruction.

3 Deviation from Additivity in Homogeneous Substitution Models.

In order to decide 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 $\frac{||D^T_\Delta, D^\star||}{w_{min}(D^\star)}$ (see Definition 2.4). This task seems hard in a general setting, but in the special case of a homogeneous substitution model it is tractable.

In the rest of this section we assume some fixed homogeneous substitution model $\mathcal{M}_{\mathbf{R}}$, defined by a unit rate matrix $\mathbf{R}$. Each evolutionary time $t > 0$ in this model is associated with a unique model matrix $\mathbf{P}(t) = e^{t\mathbf{R}}$. 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 (rather than the matrix $e^{t\mathbf{R}}$) to a dissimilarity measure $\Delta(t)$. It can be shown that such Δ is affine-additive for the model if and only if $\Delta(t) = at + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$. The *deviation* of an SR function Δ from a given affine-additive function $at + b$ in an interval $[t_0, t_1]$ is defined 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 *deviationfrom additivity* of Δ within $[t_0, t_1]$ is the minimum deviation of Δ from any affine-additive function in that interval:

Definition 3.1 (Deviation from additivity in homogeneous models)**.** *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. The* deviation from additivity of Δ in $[t_0, t_1]$ *is defined by:*

$$dev(\Delta, [t_0, t_1]) \ \stackrel{\triangle}{=} \ \inf_{a\in\mathbb{R}^+, b\in\mathbb{R}} \left\{ \ \max_{t\in[t_0,t_1]} \left\{ \frac{|\Delta(t) - at - b|}{a} \right\} \ \right\} . \tag{6}$$

Lemma 3.2 below presents a basic relation between deviation from additivity and consistency, which is used throughout this paper. In Section 4 we demonstrate the tightness of this Lemma.

Lemma 3.2. *Let $\mathcal{M}$ be a homogeneous model, and Let T be a model tree in $\mathcal{M}$ with edge lengths (measured in time units) indicated 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 $i, j \in L(T)$, let $t_{ij} = \sum_{e\in path_T(i,j)} t_e$ be the distance (sum of edge lengths) between i and j. The mapping $D_{ab}(i, j) = at_{ij} + b$ is an affine-additive mapping on $L(T)$, corresponding to edge weighting w_{ab} satisfying $w_{ab}(e) = at_e$ for each internal edge e (see the discussion following Definition 2.3). Thus we have:

$$||D_{ab}, D^T_\Delta||_\infty \ \leq \ \max_{t\in[t_0,t_1]} \{|\Delta(t) - at - b|\} \ < \ \frac{1}{2}at_{\min} \ = \ \frac{1}{2}w_{\min}(D_{ab}) \ .$$

□

Lemma 3.3 completes the picture by providing an analytic bound for $dev(\Delta, [t_0, t_1])$. For this, it assumes that Δ is a monotone increasing continuous function of t, with well-defined first and second derivatives. The lemma uses the *linear interpolation* of a the given SR function Δ in the given interval $[t_0, t_1]$, given by $\Delta_{\text{int}}(t) = At + b_0$, where $A = \frac{\Delta(t_1)-\Delta(t_0)}{t_1-t_0}$ and $b_0 = \Delta(t_0) - At_0 = \Delta(t_1) - At_1$.

Lemma 3.3. *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 $\Delta_{int}(t) = At + b_0$ be the linear interpolation of Δ in $[t_0, t_1]$ defined above, and let $B \stackrel{\triangle}{=} \max_{t\in[t_0,t_1]}\{|\Delta''(t)|\}$. Then*

$$dev(\Delta, [t_0, t_1]) \quad \leq \quad \max_{t\in[t_0,t_1]} \frac{|\Delta(t) - \Delta_{int}(t)|}{2A} \quad \leq \quad \frac{(t_1 - t_0)^2 B}{16A}. \tag{7}$$

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

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

We are looking for $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ which minimize $\frac{1}{a}\psi(a, b)$. Let $\psi_{\min} = \min_{t\in[t_0,t_1]}\{\psi(A, b_0, t)\}$, $\psi_{\max} = \max_{t\in[t_0,t_1]}\{\psi(A, b_0, t)\}$, and let $b^* = b_0 + \frac{1}{2}(\psi_{\max} + \psi_{\min})$. 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 $\psi_{\max} - \psi_{\min} \leq \frac{(t_1-t_0)^2 B}{8}$. changed - to + in def of b^*. SM

Since $\Delta_{\text{int}}(t) = At+b_0$ is a linear interpolation of Δ in $[t_0, t_1]$, we have $\psi(A, b_0, t_0) = \psi(A, b_0, t_1) = 0$. Let $t_{\min}$ be arbitrary point in the interval $[t_0, t_1]$ s.t. $\psi(A, b_0, t_{\min}) = \psi_{\min} \leq 0$ and let (t_2, t_3) be the maximal open interval in $[t_0, t_1]$ containing $t_{\min}$ in which $\psi(A, b_0, t) < 0$ (this interval can be empty if $\psi_{\min} = 0$). We define a similar interval (t_4, t_5) in which $\psi(A, b_0, t) > 0$ around some arbitrary $t_{\max}$ s.t. $\psi(A, b_0, t_{max}) = \psi_{\max}$. 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 $\psi(A, b_0, t_2) = \psi(A, b_0, t_3) = \psi(A, b_0, t_4) = \psi(A, b_0, t_5) = 0$). Therefore, the bound on the error of polynomial interpolation (see eg [16], p. 187) implies that

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

Combining these, we get

$$dev(\Delta, [t_0, t_1]) \;\leq\; \frac{1}{A}\psi(A, b^*) \;=\; \frac{1}{2A}(\psi_{\max} - \psi_{\min}) \;\leq\; \frac{\left((t_5 - t_4)^2 + (t_3 - t_2)^2\right) B}{16A} \;\leq\; \frac{(t_1 - t_0)^2 B}{16A} \;. \tag{8}$$

□

Note: In Appendix A we prove that if Δ does not intersect its linear interpolation $\Delta_{\text{int}} = At + b_0$ within the interval (t_0, t_1), then the function $At + b^*$ presented in Lemma 3.3 is 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 (8) holds in equality. The last inequality in (8) 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). These cases are important, since they hold when Δ is either convex or concave, which holds for many SR functions of interest. Rewrote the note to relax the optimality claim etc. SM

We now consider the case where the homogenous model is K2P with ti-tv ratio $R > \frac{1}{2}$. The SR function which defines the time t in the K2P model is Δ_{K2P} (since $\Delta_{\text{K2P}}(e^{t\mathbf{R}}) = t$ for all unit rate matrices $\mathbf{R}$). We compare the performance of Δ_{K2P} with that of Δ_{JC}, which is not affine-additive in this model. First, we express Δ_{JC} as a function of the ti-tv ratio R and the time t.

$$\begin{aligned}\Delta_{\mathrm{JC}}(R,t) &= -\frac{3}{4}\ln\left(1-\frac{4}{3}(p_\alpha(R,t)+2p_\beta(R,t))\right)\\ &= -\frac{3}{4}\ln\left(\frac{1}{3}e^{-\frac{2t}{R+1}}+\frac{2}{3}e^{-t\frac{2R+1}{R+1}}\right)\\ &= -\frac{3}{4}\ln\left(\frac{1}{3}e^{-\frac{2t}{R+1}}\left(1+2e^{t\frac{2R-1}{R+1}}\right)\right)\\ &= \left(\frac{3}{2(R+1)}\right)t-\frac{3}{4}\ln\left(\frac{1}{3}\left(1+2e^{-t\frac{2R-1}{R+1}}\right)\right). \qquad (9)\end{aligned}$$

Note that the homogenous K2P sub-model with $R=\frac{1}{2}$ is the JC model; in this case the second term of (9) nullifies, 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 3.3 to bound the deviation of $\Delta_{\mathrm{JC}}(R,t)$ from additivity. Denoting $\rho=\frac{2R-1}{R+1}$, we get

$$\frac{\partial\Delta_{\mathrm{JC}}(R,t)}{\partial t} = \frac{3}{2(R+1)}+\frac{3}{2}\rho\frac{e^{-\rho t}}{1+2e^{-\rho t}} > 0\ . \qquad (10)$$

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

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

We get that for any given ti-tv ratio $R>\frac{1}{2}$, $\Delta_{\mathrm{JC}}(R,t)$ 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 3.3, $dev(\Delta_{\mathrm{JC}},[t_0,t_1])=\frac{1}{2}\max_{t\in[t_0,t_1]}\{|\Delta_{\mathrm{JC}}(t)-\Delta_{\mathrm{int}}(t)|\}$ (where Δ_{int} is the linear interpolation of Δ_{JC} in $[t_0,t_1]$). A bound on this deviation from additivity can obtained through Lemma 3.3 with the following values for A and B:

$$A = \frac{\Delta_{\mathrm{JC}}(R,t_1)-\Delta_{\mathrm{JC}}(R,t_0)}{t_1-t_0}\ . \qquad (13)$$

$$B = \max_{t\in[t_0,t_1]}\left\{\left|\frac{\partial^2\Delta_{\mathrm{JC}}(R,t)}{\partial t^2}\right|\right\} = \begin{cases}-\frac{\partial^2\Delta_{\mathrm{JC}}(R,t_0)}{\partial t^2} & \text{if } \frac{\ln(2)}{\rho}<t_0\\ \frac{3}{16}\rho^2 & \text{if } \frac{\ln(2)}{\rho}\in[t_0,t_1]\\ -\frac{\partial^2\Delta_{\mathrm{JC}}(R,t_1)}{\partial t^2} & \text{if } \frac{\ln(2)}{\rho}>t_1\end{cases}\ . \qquad (14)$$

Rewrote the following paragraph. SM110703

Next, we evaluate the *stochastic noises* of Δ_{JC} and of Δ_{K2P} in a given interval $[t_0,t_1]$. Informally, the value of the random variable $\Delta_{\mathrm{K2P}}(\widehat{\mathbf{P}})$ depends on the estimation of the ti-tv ratio R from the input matrix $\widehat{\mathbf{P}}$, and hence it has a larger stochastic noise than $\Delta_{\mathrm{JC}}(\widehat{\mathbf{P}})$, which assumes that $R=\frac{1}{2}$. To put the stochastic noises of Δ_{K2P} and of Δ_{JC} in the same scale, we replace Δ_{K2P} by the affine-additive function which minimizes $dev(\Delta_{\mathrm{JC}},[t_0,t_1])$. By the note following Lemma 3.3, this function is of the form $At+b^*$, where A is given in (13) above and $b^*\in\mathbb{R}$. The stochastic noises of these functions are expressed as their *standard deviations*, $\sigma(\Delta_{\mathrm{JC}})$ and $\sigma(\Delta_{\mathrm{int}})=A\sigma(\Delta_{\mathrm{K2P}})$. We use the result in [12] to get a first order approximation (based on the delta method [21]) of $\sigma(\Delta_{\mathrm{K2P}})$ for sequences of length k and model parameters R,t:

replaced α,β in (15) by R,t. SM110627

$$\sigma(\Delta_{\mathrm{int}}) = A\sigma(\Delta_{\mathrm{K2P}}) \approx A\sqrt{\frac{(e^{\frac{4t}{R+1}}-1)+4(e^{\frac{2t}{R+1}}-1)+2(e^{\frac{4Rt}{R+1}}(e^{\frac{4t}{R+1}}+1)-2)}{16k}}\ . \qquad (15)$$

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

$$\sigma(\Delta_{\mathrm{JC}}) \approx \sqrt{\frac{p(t)(1-p(t))}{k(1-\frac{4}{3}p(t))^2}}, \tag{16}$$

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

changed α, β to R also here. SM

Figure 1 provides an illustrative comparison of the deviation-from-additivity and the stochastic-noise in the case where the ti-tv ratio is $R = 10$, and the interval $[t_0, t_1]$ is $[0.8, 2]$: Figure 1a depicts Δ_{JC} and its linear interpolation Δ_{int} in that interval. X in that figure denotes the value $\max\{|\Delta_{\mathrm{JC}}(t) - \Delta_{\mathrm{int}}(t)| : t \in [0.8, 2]\}$. Note that, by Lemma 3.3 and the subsequent note, $dev(\Delta_{\mathrm{JC}}, [0.8, 2]) = \frac{1}{2}X$. Figure 1b shows Δ_{JC} in the same setting with its closest affine-additive function $\Delta_{\mathrm{int}} + \frac{1}{2}X$. Figure 1b also shows the *stochastic* error margins of Δ_{JC} and $\Delta_{\mathrm{int}} + \frac{1}{2}X$. These stochastic error margins are defined by the above mentioned first-order approximations of the standard deviations of these functions, and are inversely proportional to the sequence length, which is taken to be 500 bp. Note how the margins of Δ_{JC} are actually more tightly concentrated around its affine-additive approximation $\Delta_{\mathrm{int}} + \frac{1}{2}X$ than the margins of that affine-additive approximation. This implies that in this setting, distances obtained by using the non-affine-additive function Δ_{JC} are actually more likely to be near-additive than distances obtained by using the additive one Δ_{K2P}. Additional demonstrations of this phenomenon are presented in the next three sections.

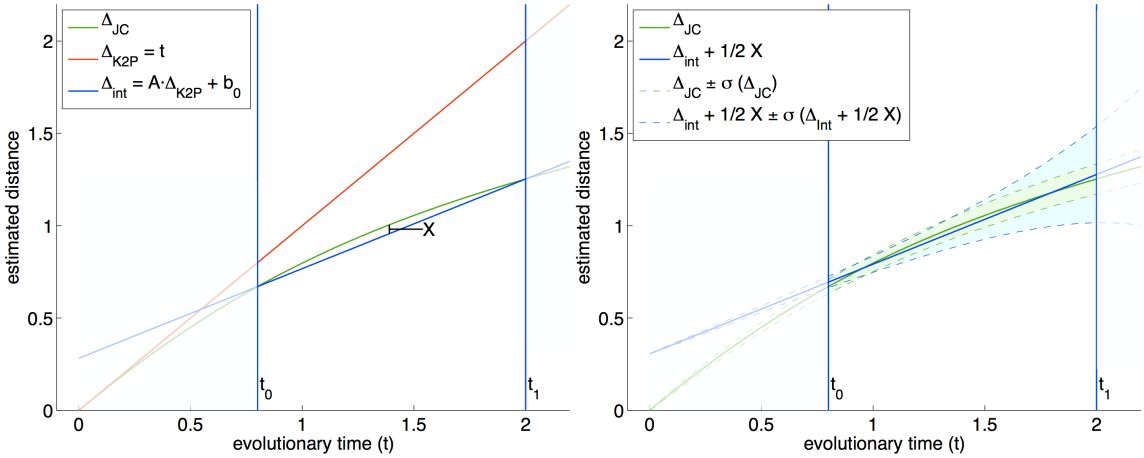


Figure 1: **Deviation from additivity and stochastic error. (a)** Δ_{JC} is portrayed (green) in the homogeneous sub-model of $\mathcal{M}_{\mathrm{K2P}}$ with $R = 10$ in the interval $t \in [0.8, 2]$. Its linear interpolation in that interval, $\Delta_{\mathrm{int}} = At + b_0$, is plotted in blue, and the maximum difference between the two functions is designated by X. The deviation of Δ_{JC} from additivity within this setting is $\frac{X}{2A}$ (A being the slope of Δ_{int}). **(b)** The affine-additive SR function minimizing its deviation from Δ_{JC} is $\Delta_{\mathrm{int}} + \frac{1}{2}X$. The stochastic error margins for the two SR functions, assuming sequence length of 500 bp, are indicated by the area between dashed lines.

4 Performance of Non affine-additive SR Functions in Quartet Resolution

In this section we demonstrate the advantage of using non-affine-additive SR functions which have small stochastic noise for quartet reconstruction. We assume that the true model is a homogeneous K2P tree with ti-tv ratio $R > \frac{1}{2}$, and compare the performance of the function $\Delta_{\rm JC}$ (which is not affine-additive) with the performance of $\Delta_{\rm K2P}$. The topology of a quartet spanning four taxa $\{1,2,3,4\}$ is represented by 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 quartet resolution algorithms essentially reduces to the four-point method (FPM) [39, 5], 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 first discuss the impact of deviation from additivity on the values of these three sums under different quartet configurations.

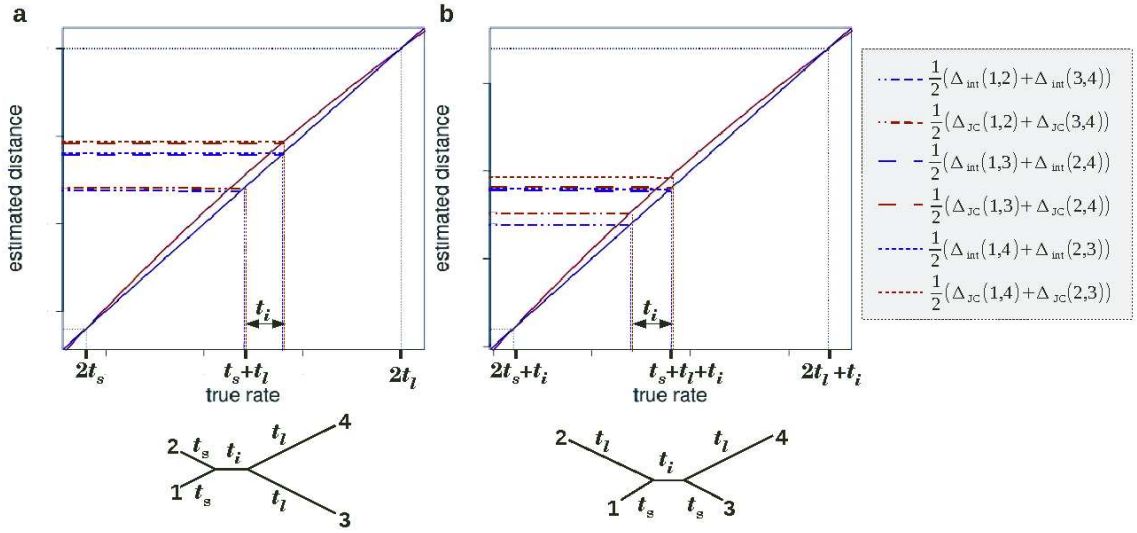


Figure 2: **Performance of the Four Point Method using $\Delta_{\rm JC}$ on K2P quartets with ti-tv ratio $R = 2$.** The concave non affine-additive SR function $\Delta_{\rm JC}$ is shown (solid red line) in the interval $[t_0, t_1]$, where t_0 and t_1 are the smallest and largest of the six pairwise distances (resp.). The solid blue line shows the linear interpolation $\Delta_{\rm int} = At + b_0$ of $\Delta_{\rm JC}$ in the interval $[t_0, t_1]$, as defined in Lemma 3.3. Horizontal bars correspond to half of each of the three sums computed by FPM under the two SR functions (see legend to the right). **(a)** In quartets of type A, $t_0 = \Delta_{\rm int}(1,2) = \Delta_{\rm JC}(1,2)$, and $t_1 = \Delta_{\rm int}(3,4) = \Delta_{\rm JC}(3,4)$, and so $\Delta_{\rm int}(1,2)+\Delta_{\rm int}(3,4) = \Delta_{\rm JC}(1,2)+\Delta_{\rm JC}(3,4)$. However, for $i \in \{1,2\}$ and $j \in \{3,4\}$, $\Delta_{\rm int}(i,j) < \Delta_{\rm JC}(i,j)$. Therefore, the deviation from additivity of $\Delta_{\rm JC}$ *increases* its FPM separation, compared to that of $\Delta_{\rm int}$. **(b)** In quartets of type B, $t_0 = \Delta_{\rm int}(1,3) = \Delta_{\rm JC}(1,3)$, and $t_1 = \Delta_{\rm int}(2,4) = \Delta_{\rm JC}(2,4)$, and so $\Delta_{\rm int}(1,3) + \Delta_{\rm int}(2,4) = \Delta_{\rm JC}(1,3) + \Delta_{\rm JC}(2,4)$. However, $\Delta_{\rm int}(1,2) = \Delta_{\rm int}(3,4) < \Delta_{\rm JC}(1,2) = \Delta_{\rm JC}(3,4)$, and so $\Delta_{\rm int}(1,2)+\Delta_{\rm int}(3,4) < \Delta_{\rm JC}(1,2)+\Delta_{\rm JC}(3,4)$. Therefore, the deviation from additivity of $\Delta_{\rm JC}$ *decreases* its FPM separation, compared to that of $\Delta_{\rm int}$.

Figure 2 demonstrates the effect of the concavity of $\Delta_{\rm JC}$ on the accuracy of the FPM on two types of quartets. Both types have an internal edge of length t_i, two long external edges of length t_l, and two short external edges of length t_s. In both types the quartet split is $(12|34)$. In type

A quartets (Fig. 2a), the short edges are on one side of the split and the long edges are on the other side. In this case, the sum associated with the split $(d_{12} + d_{34})$ is of the smallest and largest interleaf distances. The concavity of $\Delta_{\rm JC}$ increases the separation between this sum and the other two competing sums, leading to an *improvement* in reconstruction accuracy. The other quartet configuration (type B; Fig. 2b) has a short edge and a long edge on both sides of the split. In this case, the interval of inerpolation is $[d_{13}, d_{24}]$, and the distance $d_{12} = d_{34}$ is in the center of this interval. Thus the concavity of $\Delta_{\rm JC}$ decreases the separation between the sums $d_{13} + d_{24}$ and $d_{12} + d_{34}$ by approximately twice the deviation from additivity of $\Delta_{\rm JC}$ 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 $\Delta_{\rm JC}$ becomes inconsistent. Note that this demonstrates the tightness of the bound of Lemma 3.2, and in this sense, type B quartets provide a worst case scenario for quartet resolution by a concave SR function[3].

Interestingly, $\Delta_{\rm JC}$ ends up performing better than $\Delta_{\rm K2P}$ on many of its "worst case scenario" type B quartets, since its smaller stochastic noise compensates for its deviation from additivity (see also Fig. 1). This is demonstrated in the experiment described in Figure 3a, where series of homogeneous K2P quartets of type B with ti-tv ratio $R = 5$ are considered as follows: The edge lengths were set to $t_i = 0.2$, $t_l = 1.0$, and t_s varied in the interval $[0.2, 1.0]$. A total of 100,000 simulations were generated per quartet, using 1000 bp long sequences. For each simulated instance, two versions of pairwise distances were computed: one version using $\Delta_{\rm JC}$ and another version using $\Delta_{\rm K2P}$. The four-point method was invoked on both versions of the pairwise distances, and the resulting quartet split was compared to the original one. For each quartet, we recorded the number of times (out of 100,000) it was accurately resolved from either method. Despite its deviation from additivity, $\Delta_{\rm JC}$ outperforms the additive SR function $\Delta_{\rm K2P}$ on many of these quartets. Only when the deviation from additivity is sufficiently large ($t_l/t_s > 3.6$ in these experiment), $\Delta_{\rm K2P}$ outperforms $\Delta_{\rm JC}$.

changed the analysis here. SM

[3]Types A and B quartets correspond to *Farris zones* and *Felsenstein zones* (resp.) - see eg [8], Chapter 9.

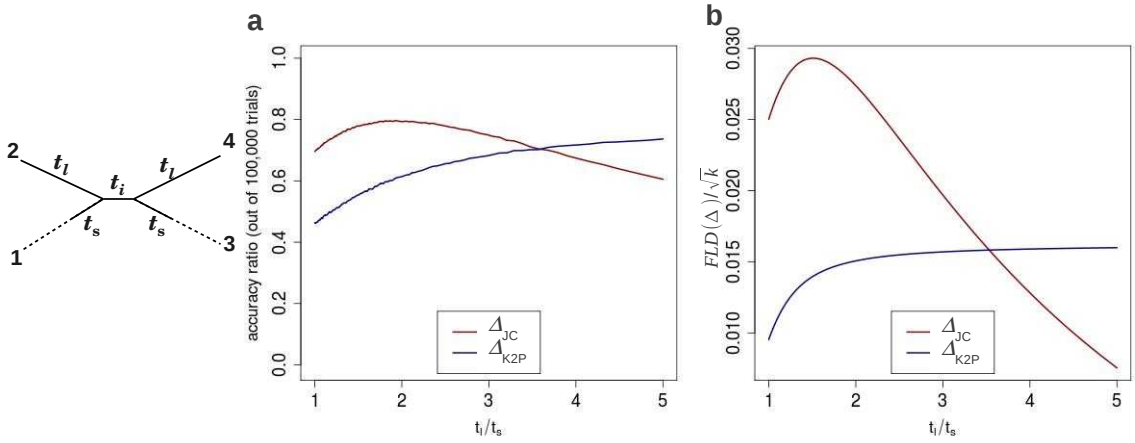


Figure 3: **Performance of $\Delta_{\rm JC}$ and $\Delta_{\rm K2P}$ on a series of quartets of type B.** A series of homogeneous K2P quartets is considered (left illustration), with ti-tv ratio of $R = 5$, and edge lengths $t_i = 0.2$, $t_l = 1$, and $t_s \in [0.2, 1]$. **(a)** Reconstruction accuracy using FPM and either $\Delta_{\rm JC}$ (red) or $\Delta_{\rm K2P}$ (blue) plotted against t_l/t_s. Accuracy ratio is estimated using 100,000 independent replicates for each parameter setting and sequences of length 1000 bp. **(b)** Fisher's Linear Discriminant (FLD) for the sums corresponding to splits (12|34) and (13|24) under either $\Delta_{\rm JC}$ (red) or $\Delta_{\rm K2P}$ (blue) plotted against t_l/t_s.

The trends observed in Figure 3a can be explained as follows. As t_s shrinks, the minimal estimated distance decreases and the maximal estimated distance is unchanged, hence the constant B in Lemma 3.3 cannot decrease. Hence the bound of Lemma 3.3 on the deviation from additivity of $\Delta_{\rm JC}$ is at least proportional to $\frac{(t_1-t_0)^3}{\Delta_{\rm JC}(t_1)-\Delta_{\rm JC}(t_0)}$. This last value increases as t_s decreases, in correlation with the deterioration of the performance of $\Delta_{\rm JC}$ relative to that of $\Delta_{\rm K2P}$.

4.1 Using Fisher's Linear Discriminant.

In order to provide a better understanding (and ability to predict) the results of similar experiments, we present a simple and general framework based on Fisher's linear discriminant (FLD). FLD measures the separation between normal random variables $X \sim N(\mu_1, \sigma_1)$ and $Y \sim N(\mu_2, \sigma_2)$ using the following measure ([10, 2]):

$$FLD(X,Y) \;=\; \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \;. \tag{17}$$

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

"We note that..." requires some expert validation..SM

We note that FLD requires that the random variables X, Y are independently distributed normal random variables, and this is not the case here: The three sums are not normally distributed, neither are independent (since they are correlated through the substitution process along the external edge). Nonetheless, as Figure 3 shows, FLD provides a quite reliable comparison of the expected performance of $\Delta_{\rm JC}$ and $\Delta_{\rm K2P}$ on the given quartets. Figure 3b plots FLD of $\Delta_{\rm JC}$ and $\Delta_{\rm K2P}$ associated with the comparison of the true split (12|34) and the "$\Delta_{\rm JC}$ favored split" (13|24) along the quartet series considered in Figure 3a. As shown, the equilibrium point of the Fisher discriminants of $\Delta_{\rm JC}$ and $\Delta_{\rm K2P}$ is pretty close to the equilibrium point of the accuracy of reconstructions of these two functions.

Perhaps the most useful feature of this framework is the natural way in which it teases apart the stochastic error from the deviation from additivity. If we denote the enumerator of FLD by SEP and its denominator by $NOISE$, then a comparison of FLD estimates between two SR function Δ_1, Δ_2 can be represented as a ratio of ratios:

$$\frac{FLD(\Delta_1)}{FLD(\Delta_2)} \;=\; \frac{SEP(\Delta_1)}{SEP(\Delta_2)} \;/\; \frac{NOISE(\Delta_1)}{NOISE(\Delta_2)} \;. \tag{18}$$

Comparison of the SEP and $NOISE$ ratios for $\Delta_{\rm JC}$ and $\Delta_{\rm K2P}$ is demonstrated for four series of homogeneous K2P quartet in Figure 4. The bottom-left series is the same one considered in Figure 3. The main trends shown in this figure are quite the expected ones: (a) the $NOISE$ ratio favors $\Delta_{\rm JC}$ as the maximal distance increases (hence it is almost constant at the left two plots, where the maximal distance is fixed, and monotone decreasing at the right ones): the larger is the effect of the noise, the more it favors $\Delta_{\rm JC}$; (b) the $NOISE$ ratio also favors $\Delta_{\rm JC}$ as the ti-tv ratio increases, since $\Delta_{\rm K2P}$ becomes more noisy but $\Delta_{\rm JC}$ is less affected (this is why the $NOISE$ ratio for $R = 5$ is consistently smaller than for $R = 2$); (c) the SEP ratio favors $\Delta_{\rm K2P}$ when the quartet becomes unbalanced, since the gap between the minimal and maximal distances increases, and the factor $(t_1 - t_0)^2$ in the bound of Lemma 3.3 becomes larger (this is why the SEP curve is decreasing in all four sub-figures); (d) the SEP ratio favors $\Delta_{\rm K2P}$ also as the ti-tv ratio increases, since the deviation from additivity of $\Delta_{\rm JC}$ increases, ie the implied value of B in the bound of Lemma 3.3 becomes larger (this is why the SEP curves for $R = 5$ are lower than for $R = 2$). Analysis of this type, using FLD to predict relative accuracy of quartet reconstruction, are likely to be easily extended to more complex homogeneous models and SR functions.

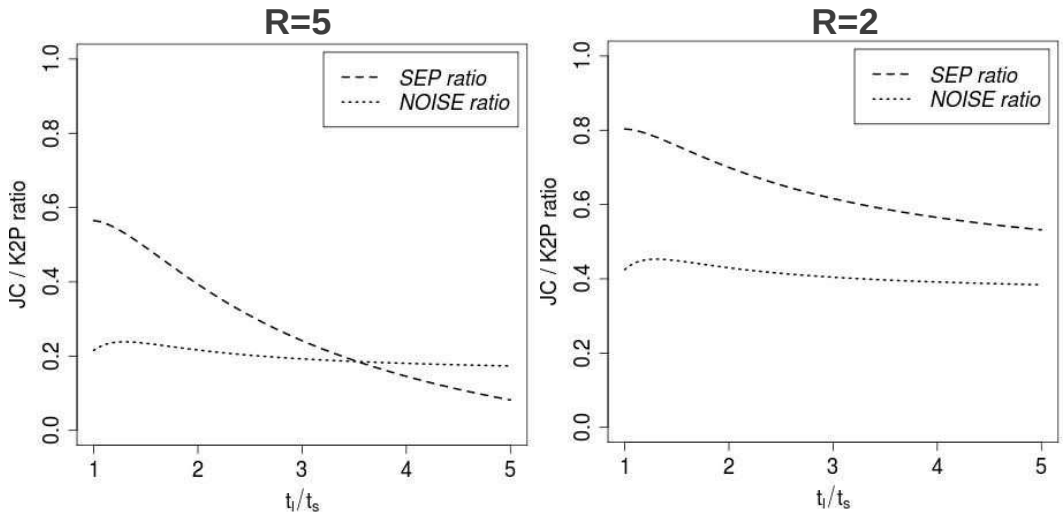


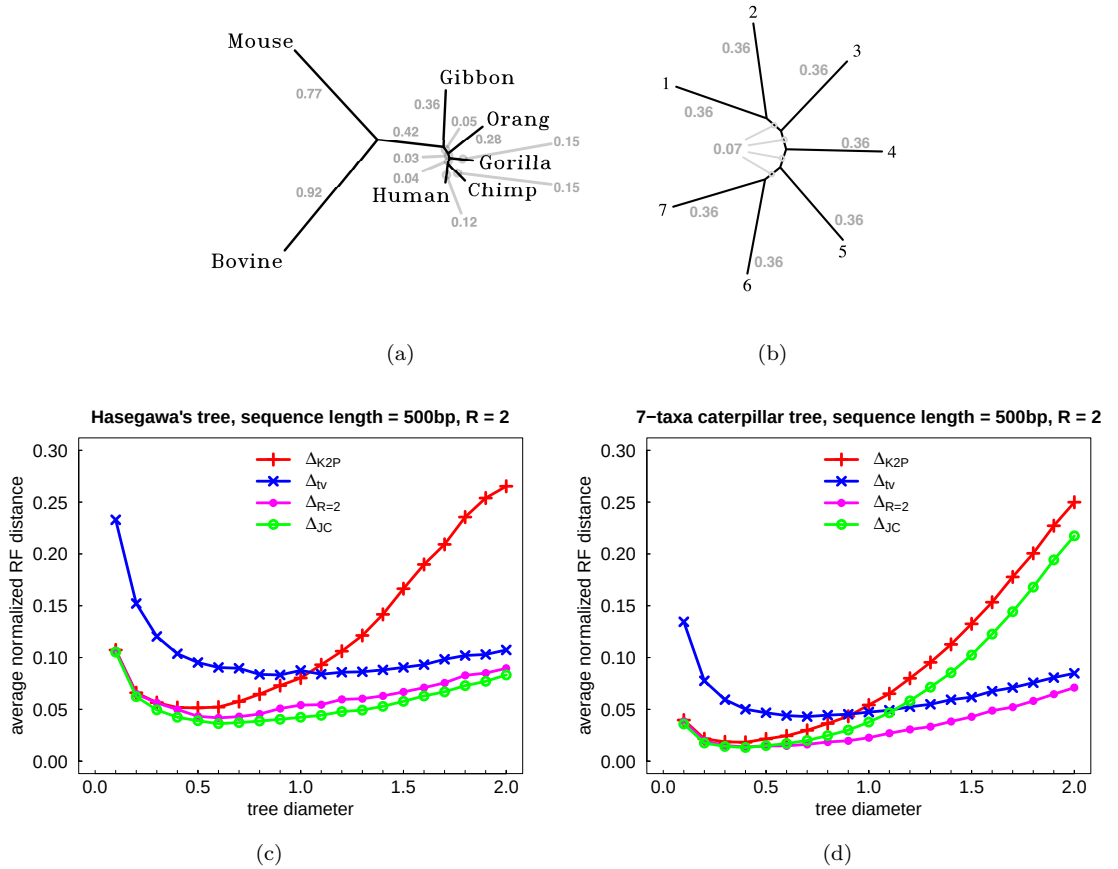
Figure 4: SEP **and** $NOISE$ **ratios.** $SEP(\Delta_{\mathrm{JC}})/SEP(\Delta_{\mathrm{K2P}})$ (dashed) and $NOISE(\Delta_{\mathrm{JC}})/NOISE(\Delta_{\mathrm{K2P}})$ (dotted) plotted against t_l/t_s for four series of homogeneous K2P quartets of type B. Top two series have ti-tv ratio of $R = 2$, 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.

5 Simulations on Hasegawa's Tree

In this section we explore the effects of deviation from additivity and stochastic noise on the tree assembled by Hasegawa, Kishino, and Yano in 1985 [15, 8], which spans seven eutherian mammals, and whose original reconstruction was based on mitochondrial DNA sequences. This tree (Fig. 5(a)) has a caterpillar topology (meaning that every internal node touches 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 use the tree structure and edge lengths to generate simulated data sets. We consider 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 [26, 32] to these distance matrices and recording the *Robinson-Foulds* (RF) [24] distance between the reconstructed tree and the Hasegawa tree. Sequence simulation was performed using SeqGen [23] (by choosing the HKY model with uniform base frequencies), and tree reconstruction was performed using the version of NJ implemented in the PHYLIP package [7].

In our results, we compare four different SR functions: Δ_{JC}, Δ_{K2P}, Δ_{tv}, and $\Delta_{\mathrm{R=2}}$. The first two are as described in Equations (5) and (4), respectively. The third SR function, Δ_{tv}, considers



(a) (b)

(c) (d)

Figure 5: **Simulations on Hasegawa's Tree.** **(a)** Hasegawa's tree spanning seven eutherian mammals [15]. **(b)** A semi-symmetric caterpillar tree spanning seven terminal taxa with uniform internal edge lengths t_{int} and uniform external edge lengths $t_{ext} = 5t_{int}$. **(c)** Reconstruction accuracy of four different SR functions on different scaled versions of Hasegawa's tree. **(d)** Reconstruction accuracy of four different SR functions on different scaled versions of the semi-symmetric caterpillar tree. In both plots, the scale of the tree (X axis) is measured by its diameter, and reconstruction accuracy (Y axis) is measured by the average normalized RF distance between the reconstructed tree and the true tree.

only tv-type substitutions: $\Delta_{\rm tv}(p_\beta, t) = -\frac{1}{4}\log(1-4p_\beta(t)) = \beta t$, and the fourth SR function, $\Delta_{\rm R=2}$, is based on a maximum likelihood (ML) estimator of the time $t = \alpha t + 2\beta t$ from the estimated values $\hat{p}_\alpha(t), \hat{p}_\beta(t)$, given that $R = 2$. Informally, this function is not more noisy than $\Delta_{\rm JC}$, and it is also additive since it assumes the correct model parameters.

added some more on $\Delta_{\rm R=2}$ SM110630

The performance of these four SR functions is traced across the different tree scales in Figure 5(c). For each SR function Δ and scale s, we record the average normalized RF distance of trees reconstructed using Δ from the 10,000 data sets generated under scale s. RF distance is 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 [12], we see that $\Delta_{\rm K2P}$ performs well in shorter scales, and $\Delta_{\rm tv}$ performs well in longer scales. However, both additive SR functions are significantly outperformed in nearly all cases by $\Delta_{\rm JC}$. Surprisingly, $\Delta_{\rm JC}$ even slightly outperforms $\Delta_{\rm R=2}$, which is additive in this case (since the simulated ti-tv ration is $R = 2$) and is supposed to have minimal stochastic error. We hypothesize that this happens since Hasegawa's tree structure posses a bias whose nature is similar

to that of type A quartets, which improves the performance of concave SR functions such as Δ_{JC}, as we discussed in Section 4.

To test this hypothesis, we went through a similar experiment with a more symmetric seven-taxon caterpillar tree, with internal edges of uniform length t_{int}, and external edges of uniform length $t_{ext} = 5t_{int}$(Fig. 5(b);5(d)). The three additive SR functions show similar performance trends in both trees. However, Δ_{JC} performs much poorly on the second caterpillar tree, presumably, since deviation from additivity is inhibiting its accuracy in this case. Despite this fact, Δ_{JC} still outperforms Δ_{K2P} in all scales and Δ_{tv} in the smaller scales ($s < 1.1$). Clearly, when the ti-tv ratio is known (say, $R = r$), the SR function $\Delta_{\mathrm{R=r}}$ is the optimal choice for distance computation, since it combines additivity with reduced stochastic noise. However, these experiments provide additional evidence for the usefulness of Δ_{JC} in reconstructing homogeneous K2P trees with unkown ti-tv ratio.

6 Inferring Trees from Genomic Sequences

In this section we explore the performance of three SR functions when reconstructing trees from genomic DNA sequences. Next to Δ_{JC} and Δ_{K2P} we also computed distances using the well kown LogDet SR function [30, 20], denoted here as Δ_{LogDet}. Extending our discussion to this setting is challenging in two respects. First of all, unlike in 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 unkown 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 Δ_{LogDet} is additive). Hence, we have to assume in this case that Δ_{JC}, Δ_{K2P}, and Δ_{LogDet} are non affine-additive, where Δ_{JC} and Δ_{K2P} are still likely to exhibit higher deviation from additivity than Δ_{LogDet}, since they make stronger assumptions on the substitution model.

6.1 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 between a large number of species in [3, 37]. These 31 gene families were selected to capture the evolutionary history of the species from which they are extracted. This was done in [3] 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 [3, 37] to obtain reliable multiple-sequence alignements for each COG[4]. We computed a 199-way multiple alignment of the protein sequences of each COG using HMMalign [4] 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 [33] to filter out alignment columns with 50% or more gap symbols. The alignments were manually scanned, and 36 species which contribute a large number of gaps to the alignments were removed from 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 Living Tree Project [38]. This tree, spanning 8,029 species at the time of writing, is based on 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.

[4]Our procedure differs from that of [3, 37] in that we have to convert the alignments to DNA sequence alignments.

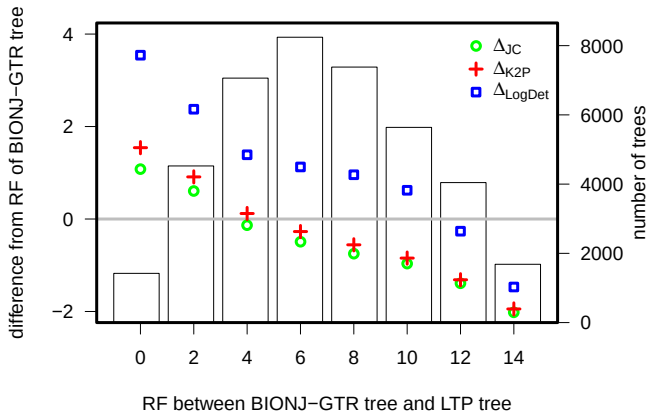


Figure 6: **Evaluation against PhyML Tree.** The 40,000 subsets of size 10 were partitioned according to the the RF distance of the tree reconstructed by PhyML from the LTP tree (X axis). The Y axis describes the difference between the RF distance associated with a particular SR function (Δ_{K2P}, Δ_{JC}, or Δ_{LogDet}) and the RF distance associated with PhyML. The bar plot in the background depicts the proportional number of subsets in each partition.

6.2 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 matricess – 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.

In order to study trends among the 40,000 subsets, we attempted to sort them according to "hardness of reconstruction" by a fourth independent reconstruction technique. For this, we applied PhyML [14], using the BIONJ reconstruction algorithm [11] on distances obtained under the general time-reversible model with invariant sites and Gamma distribution of rates across sites (GTR+Γ+I) [19, 25]. The GTR+Γ+I model is a highly general substitution model, which is expected to provide an increased fit to the sequence data. Consequently, Δ_{LogDet} should feature a much lower deviation from additivity under the assumed model since it is additive in the GTR model. The 40,000 sampled subsets of the alignment were partitioned according to the RF distance between the PhyML tree and the true (LTP) tree. Subsets corresponding to low PhyML RF distance are considered to be easier to reconstruct compared to subsets corresponding to high PhyML RF distance.

slightly changed here in response to REVIEW 1 comment SM110630

Results are shown in Figure 6. Of the 40,000 trees inferred under Δ_{JC}, 83.1% show an equal or lower RF distance to corresponding subsets of the LTP tree than those reconstructed by PhyML under the GTR+Γ+I model. Moreover, Δ_{JC} outperforms Δ_{K2P} and Δ_{LogDet} on average in all partitions, and Δ_{LogDet} shows by far the worst performance with 48.7% of all reconstructed trees

achieving higher RF distances to the reference trees than those inferred by PhyML. This result endorses our line of argument, since the SR functions with lower stochastic error but inferior fit perform best. Thus the deviation from additivity caused by by choosing models that deviate further from the assumed underlying model, is small compared to the gain in accuracy achieved by the reduced variance in distance estimation.

7 Discussion and Outlook

In this paper we study basic properties of evolutionary distance estimation using SR functions and how they affect the accuracy of phylogenetic reconstruction. When studying accuracy of statistical estimates, it is important to consider both the bias of the estimate and its variance (referred to as stochastic noise). In some cases it might be worth trading off variance for bias, resulting in a slightly skewed estimate which is less noisy. A challenge in carrying out such a study for statistical estimation of evolutionary distances is that bias is not completely well-defined in this case, since the "true" evolutionary distances can take many forms; any affine-additive SR function is a valid one for the purpose of phylogenetic reconstruction.

We introduce the concept of *deviation from additivity* to quantify the bias of an SR function in a homogeneous substitution model. We demonstrate this analytic framework by studying the bias of the Jukes-Cantor SR function (Δ_{JC}) in Kimura's two parameter model when the ti-tv ratio is significantly larger than $\frac{1}{2}$. We show that even when the ti-tv ratio is as high as 5 or 10, this bias is small enough such that the reduced variance of Δ_{JC} makes it overall more accurate than Kimura's SR function (Δ_{K2P}), which has no bias. We show this using analytic bounds as well as detailed simulation experiments on quartet trees. We also introduce a useful, simple and general heuristic, based on the Fisher's linear discriminant (FLD), for predicting scenarios in which a simplified, non affine-additive function is likely to perform better than an additive one. added here SM110628

Experiments on simulated data, simulating evolution along the Hasegawa's tree, show that for this specific tree, the deviation from additivity increases the reconstruction probability even w.r.t. an additive SR function which is not more noisy than the non-affine-additive one. Finally, our results were also affirmed in a round of experiments on real biological sequences. In the case of real data, the true substitution model is likely to be very complex, and all common distance formulas are expected to have some bias. Our results show that simpler SR functions with lower variance lead to more accurately reconstructed trees on average, compared to SR functions that are expected to have reduced bias but higher variance.

With the devised framework at hand, the study of distance estimation can be extended in different directions. More complex models and non-additive SR functions could be studied, and improved methods for the analysis of biological sequences could be established. Additionally, there is a need for extending the FLD-based heuristic to trees larger than quartets. Finally, incorporating our methods in existing software for phylogenetic reconstruction looks like a promising venue for increasing the accuracy of distance-based phylogenetic reconstruction at low (or even negative) computational cost.

References

[1] K. Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.

[2] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–1287, March 2006.

[4] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1999.

[5] P. Erdos, M. Steel, L. Szekely, and T. Warnow. A few logs suffice to build (almost) all trees (I). *Random Structures Algorithms*, 14:153–184, 1999.

[6] W. J. Ewens and G. Grant. *Statistical Methods in Bioinformatics: An Introduction*. Springer, 2005.

[7] J. Felsenstein. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989.

[8] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associated, Inc., Sunderland, MA, 2004.

[9] J. Felstenstein and E. Sober. Parsimony and likelihood: an exchange. *Systematic Zoology*, 35:617–626, 1986.

[10] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:177–188, 1936.

[11] O Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*, 14(7):685–695, 1997.

[12] I. Gronau, S. Moran, and I. Yavneh. Towards optimal distance functions for stochastic substitution models. *J Theor Biol*, 260(2):294–307, 2009.

[13] I. Gronau, S. Moran, and I. Yavneh. Adaptive distance measures for resolving K2P quartets: Metric separation versus stochastic noise. *J Comp Biol*, 17(11):1391–1400, 2010.

[14] S. Guindon and O. Gascuel. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52:696–704, 2003.

[15] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, October 1985.

[16] L.W. Johnson and R.D. Riess. *Numerical Analysis*. Addison Wesley, 1977.

[17] T. Jukes and C. Cantor. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[18] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, December 1980.

[19] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20:86–93, 1984.

[20] P. Lockhart, M. Steel, M. Hendy, and D. Penny. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*, 11(4):605–612, 1994.

[21] G. Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

[22] A. Papoulis and S. U. Pillali. *Probability, Random Variables and Stochastic Processes*. McGraw Hill, 4th edition, 2002.

[23] Andrew Rambaut and Nicholas C. Grass. Seq-Gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences*, 13(3):235–238, June 1997.

[24] F. Robinson and R. Foulds. Comparison of phylogenetic trees. *Math Biosci*, 53:131–147, 1981.

[25] F. Rodriguez, J. L. Oliver, A. Marin, and J. R. Medina. The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, 142:485–501, Feb 1990.

[26] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4:406–425, 1987.

[27] S. Sattath and A. Tversky. Additive similarity trees. *Psychometrica*, 42(3):319–345, 1977.

[28] C Semple and M Steel. *Phylogenetics*. Oxford, 2003.

[29] Elliot Sober. A likelihood justification of parsimony. *Cladistics*, 1:209–233, 1985.

[30] M. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl Math Lett*, 7(2):19–24, march 1994.

[31] M. Steel and D. Penny. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.*, 17:839–850, Jun 2000.

[32] J. Studier and K. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol*, 5(6):729–731, 1988.

[33] G. Talavera and J. Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, 56:564–577, Aug 2007.

[34] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10(3):512–526, May 1993.

[35] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol*, May 2011.

[36] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[37] C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126–1130, February 2007.

[38] P. Yarza, W. Ludwig, J. Euzeby, R. Amann, K. H. Schleifer, F. O. Glockner, and R. Rossello-Mora. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Syst. Appl. Microbiol.*, 33:291–299, Oct 2010.

[39] K. Zaretskii. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Mat Nauk*, 20(6):90–92, 1965. in Russian.

A Tightness of Lemma 3.3.

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 for some b^*, the linear function closest to f in that interval under the L^∞ norm is $At + b^*$ (f represents the function Δ in Lemma 3.3). We use the following notations, conforming to the notations in the proof of Lemma 3.3:

tried to make the opening sentence more "general" SM110630

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

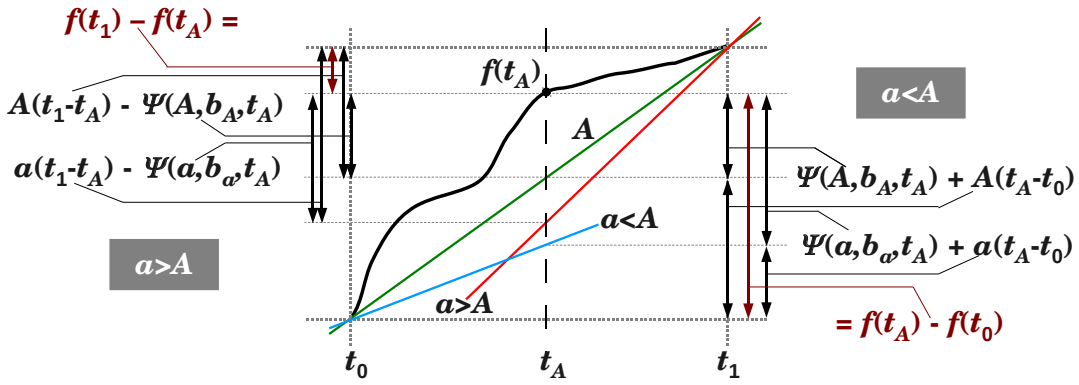


Figure 7: **Proof of Lemma A.1.** A function $f(t)$ is portrayed (bold) with its linear interpolation $At + b_A$ (green) in the interval $[t_0, t_1]$, s.t. $f(t) \geq At + b_A$ for all $t \in [t_0, t_1]$. Equation (19) is illustrated for $a < A$ on the right, and equation (20) is illustrated for $a > A$ on the left.

Lemma A.1. *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 when $f(t) \geq At + b$ for all $t \in [t_0, t_1]$. The other case (when $f(t) \leq At + b$ 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 $t \in [t_0, t_1]$. It is not difficult to then see that $\psi(a) = \frac{1}{2}\psi(a, b_a)$. If A is the slope of the linear interpolation of $f(t)$ in $[t_0, t_1]$, then the offset of that interpolation is given by b_A. 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 Fig. 7).

For $a > A$, we get the following equality (Fig. 7; right):

$$\psi(A, b_A, t_A) + A(t_A - t_0) \;=\; f(t_A) - f(t_0) \;=\; \psi(a, b_a, t_A) + a(t_A - t_0)\ . \tag{19}$$

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

$$a\psi(A, b_A, t_A) + aA(t_A - t_0) \;<\; A\psi(a, b_a, t_A) + Aa(t_A - t_0) \quad \Rightarrow \quad a\psi(A, b) < A\psi(a, b')\ .$$

Similaly, if $a > A$, we get the following equality (Fig. 7; left)

$$A(t_1 - t_A) - \psi(A, b_A, t_A) \;=\; f(t_1) - f(t_A) \;=\; a(t_1 - t_A) - \psi(a, b_a, t_A)\ , \tag{20}$$

and $a > A$ implies that

$$aA(t_1 - t_A) - a\psi(A, b) > Aa(t_1 - t_A) - a\psi(a, b') \quad \Rightarrow \quad a\psi(A, b) < A\psi(a, b')\ .$$

□