A signal-to-noise analysis of phylogeny estimation by Neighbor-Joining: Insufficiency of polynomial length sequences

Michelle R. Lacey [a,*], Joseph T. Chang [b]

[a] *Tulane University, Department of Mathematics, New Orleans, LA 70118*

[b] *Yale University, Department of Statistics, New Haven, CT 06511*

Abstract

Phylogeny reconstruction is the process of inferring evolutionary relationships from molecular sequences, and methods that are expected to accurately reconstruct trees from sequences of reasonable length are highly desirable. To formalize this concept, the property of *fast-convergence* has been introduced to describe phylogeny reconstruction methods that, with high probability, recover the true tree from sequences that grow polynomially in the number of taxa n. While provably fast-converging methods have been developed, the Neighbor-Joining (NJ) algorithm of Saitou and Nei remains one of the most popular methods used in practice. This algorithm is known to converge for sequences that are exponential in n, but no lower bound for its convergence rate has been established. To address this theoretical question, we analyze the performance of the NJ algorithm on a type of phylogeny known as a "caterpillar tree." We find that, for sequences of polynomial length in the number of taxa n, the variability of the NJ criterion is sufficiently high that the algorithm is likely to fail even in the first step of the phylogeny reconstruction process, regardless of the degree of polynomial considered. This result demonstrates that, for general n-taxa trees, the exponential bound cannot be improved.

Key words: phylogeny reconstruction, distance methods, Neighbor-Joining, sequence lengths, fast-converging

* Corresponding author. Address: Department of Mathematics, 6823 St. Charles Ave., New Orleans, LA 70118-5698. Phone: 504.862.3439, Fax: 504.865.5036
Email address: mlacey@math.tulane.edu (Michelle R. Lacey).

1 Introduction

Phylogeneticists employ evolutionary models to interpret observed molecular sequence data as a series of divergences from unknown ancestral sequences. A variety of methods and algorithms have been developed to implement these models, and researchers are left with a wealth of options in choosing the phylogenetic method that best suits their needs with respect to modelling complexity, computational efficiency, and interpretability. There are advantages and disadvantages to each of the major classes of methods, and no particular algorithm has emerged as a clear "winner" in the phylogenetic research community. However, there are certain properties that are considered to be important attributes for any phylogeny reconstruction method: computational efficiency, consistency, and robustness. And, because biologists are often limited in the amount of sequence data that is available for their studies, methods that are expected to reconstruct trees accurately from sequences of practical length are highly desirable.

The distance-based Neighbor-Joining (NJ) algorithm, introduced by Saitou and Nei in 1987 [1], offers an intuitive, computationally efficient process for phylogeny reconstruction. It is easy to implement and runs quickly on large datasets, making it a popular choice for practicing biologists. In addition, the method is known to be consistent, meaning that, in the limit as the observed sequence length tends to infinity, the NJ criterion will always be minimized for a pair of neighboring leaves (see Durbin *et. al* [2] for a proof of this theorem). Atteson [3] has established conditions under which NJ will perform well from a matrix of estimated distances: if the difference between the true and estimated distances is bounded by half the length of the shortest edge in a tree T, then NJ

will correctly reconstruct the topology of T. But how long must sequences be to guarantee that the estimated distances will, with high probability, meet this criterion? Atteson also briefly addressed this question, deriving a bound that guarantees the accurate reconstruction of general n-taxa trees from sequences that are of exponential length in the maximum distance between any two leaves in the tree.

In 1999, Huson *et al.* [4] introduced a new standard for evaluating the convergence rates of phylogeny reconstruction algorithms. The authors defined a method to be *fast-converging* under a model of evolution if the method could, with high probability, accurately recover the topology of any model tree from sequences that grow only polynomially in the number of leaves. Huson *et al.* described an approach for creating fast-converging algorithms from existing distance methods known as the *Disk-Covering Method* (DCM) [4, 5, 6], proving that "DCM-boosted" methods were fast-converging under the Jukes-Cantor model. Additional research refined DCM by developing methods which meet a more general criterion known as *absolute fast-convergence* [7, 8].

After introducing these fast-converging methods, the researchers performed simulations to evaluate the performance of DCM-boosted algorithms relative to NJ and other popular methods. While finding that the provably fast-converging methods out-performed NJ for some trees, Nakhleh *et al.* [8] also reported that, in some cases, NJ offered significant improvements in accuracy over fast-converging methods, and other studies reported minimal differences between NJ and fast-converging methods on large subsets of the tree space [9, 4, 10]. While none of these simulation studies evaluated sufficiently long sequences to realistically compare polynomial and exponential convergence rates, these experimental results led to the suggestion that Atteson's

exponential convergence bound for NJ was "probably loose" [4]. However, because no immediate extension of the proof of fast-convergence of DCM-boosted methods was applicable to NJ, the question remained: Is the original Neighbor-Joining algorithm a provably fast-converging method?

To address this issue, we investigate the performance of the NJ algorithm for sequences of polynomial length, analyzing the asymptotic behavior of the method on a phylogeny known as a "caterpillar tree." Our approach focuses on the first step of the caterpillar reconstruction process, comparing the variability of the NJ criterion to its expected value. We find that this signal-to-noise ratio converges to 0, demonstrating that polynomial length sequences are insufficient to guarantee accurate performance of the NJ algorithm. In Section 2 we provide necessary background, in Section 3 we summarize our findings, and we close the paper with a discussion of the theoretical and practical implications of this analysis in Section 4. Detailed proofs of the Theorems in Section 3 are provided in the Appendix.

2 Background

We begin with necessary definitions and background to motivate and support our analysis. We provide a brief overview of distance-based methods and the Jukes-Cantor model for sequence evolution, and then review the NJ algorithm. Finally, we introduce the "caterpillar tree" and discuss our approach for analyzing the performance of the NJ method using this model tree.

2.1 Distance-based methods and the Jukes-Cantor Model

The term "distance-based" describes a class of methods that reconstruct phylogenies from a matrix of pairwise distances d_{ij}, where d_{ij} denotes the distance between leaves i and j. Phylogenetic researchers employing distance-based methods must assume that the matrix of pairwise distance estimates provides sufficient information for the accurate reconstruction of the evolutionary relationships among the taxa considered. This is a strong assumption, and it is clear that the choice of distance estimation method will have a considerable impact on the resulting phylogeny. There are several methods for estimating distances, some of which allow for the differential weighting of certain types of substitutions in general DNA sequences [11, 12], and others that specifically model the evolution of protein coding regions [13, 14]. The appropriateness of various assumptions has been evaluated using some known or experimentally generated phylogenies [15, 16, 17], and several statistical procedures have been introduced for choosing among evolutionary models in practice (see Posada and Buckley [18] for a recent overview of these methods). Most theoretical analyses, however, focus on the Jukes-Cantor [19] model. The central assumptions of the model are as follows:

(i) The equilibrium frequency of each of the four nucleotides $\{A, C, G, T\}$ is equal to $\frac{1}{4}$.

(ii) Every type of substitution is equally likely.

(iii) The rate of substitution does not vary by time or position.

(iv) Mutations at each position occur independent of mutations at every other position.

With these assumptions, the model is fully described by the following instantaneous rate matrix R:

	A	**C**	**G**	**T**
A	$-3r$	r	r	r
C	r	$-3r$	r	r
G	r	r	$-3r$	r
T	r	r	r	$-3r$

For a random variable X representing a nucleotide evolving according to the Jukes-Cantor model, we have, for any specified time t,

$$p_t = P(X_t \neq X_0) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}rt}\right). \tag{1}$$

For any pair of sequences S_i and S_j of length L, the proportion of positions at which the two sequences differ is given by the Hamming distance

$$\sum_{l=1}^{L} \frac{(S_{i,l} \neq S_{j,l})}{L}.$$

Under the Jukes-Cantor model, this proportion provides an estimate for the expected Hamming distance $p = p_t$ for any position along the sequence. Interpreting the product of substitution rate and time as a measure of "distance" (that is, $d = rt$) and inverting Equation 1, we have

$$d = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right). \tag{2}$$

As p approaches the value of $\frac{3}{4}$ (the expected probability of observing a difference between completely unrelated sequences at any position), the distance d between the two sequences becomes infinitely large.

While most short distances can be easily estimated under this model, it is clear that, for any pair of sequences with estimated Hamming distance $\hat{p} \geq \frac{3}{4}$, the Jukes-Cantor distance estimate $\hat{d}$ will be undefined. This issue presents practical problems for researchers attempting to construct distance estimates from limited amounts of DNA, since most distance-based phylogeny reconstruction algorithms will fail to produce a tree when even a single pairwise distance estimate is undefined for a set of sequences. While one would ideally avoid this problem by acquiring enough sequence data to compute extremely precise distance estimates, sufficiently long sequences are often not available for distantly related taxa. For this reason, undefined Jukes-Cantor distances are typically "corrected" and assigned a large value. This value is generally at least as large as the maximum well-defined distance estimate observed for the set of sequences, and may be determined either as a function of the observed estimates (known as a "fix factor" [20]) or simply arbitrarily assigned [21].

2.2 The Neighbor-Joining Algorithm

The Neighbor-Joining (NJ) algorithm reconstructs a phylogeny from n sequences by iteratively joining the pairs of leaves i and j which minimize the criterion

$$D_{ij} = d_{ij} - \frac{1}{n-2}\left(\sum_{k=1}^{n} d_{ik} + \sum_{k=1}^{n} d_{jk}\right),$$

where d_{ij} is the distance between sequences i and j. Once a pair of leaves is selected to join, a node m connecting this pair is added to the tree. For all nodes $k \neq \{i, j\}$, the distance d_{mk} is then defined to be

$$d_{mk} = \frac{1}{2}\left(d_{ik} + d_{jk} - d_{ij}\right),$$

and the distances d_{im} and d_{jm} are given by

$$d_{im} = \frac{1}{2}\left(d_{ij} + \frac{1}{n-2}\left(\sum_{k=1}^{n} d_{ik} - \sum_{k=1}^{n} d_{jk}\right)\right) \text{ and } d_{jm} = d_{ij} - d_{im}.$$

The leaves i and j are subsequently removed from the distance matrix and replaced with the new node m, and the NJ criterion matrix D is recomputed for the new set of $n-1$ leaf nodes. The algorithm continues to join pairs of leaves until only two leaves i and j remain, and these are connected with edge length d_{ij} to complete the phylogeny.

2.3 The Caterpillar Tree

The term "caterpillar tree" is used to describe a phylogeny in which n taxa are connected to a single spine (see Figure 1.a). If we consider a simplified "legless" caterpillar in which the n taxa are connected in sequence by edges of equal length d_e, then the distance between a pair of taxa i and j on a caterpillar tree is simply given by $|j-i|d_e$, the number of edges separating the pair (see Figure 1.b). Since the longest pairwise distance $d_{1,n} = (n-1)d_e$, Atteson's bound would require sequence lengths to be exponential in n to guarantee asympototic convergence. For this reason, the caterpillar tree provides a useful model for exploring the performance of the NJ method for sequences of polynomial length.

2.4 Method for analyzing the asymptotic stability of NJ

It is clear from the structure of the caterpillar tree that there are only two pairs of leaves that are separated by a single node, leaves 1 and 2 and leaves $n-1$

and n. Thus, the only way that the NJ algorithm can correctly reconstruct the caterpillar is by joining either of these two pairs of leaves on the first step of the process. On subsequent steps, the algorithm must continue to work its way in towards the center of the caterpillar until the tree is fully reconstructed. If, at any time, a pair of non-neighboring leaves are joined, then the NJ algorithm will fail. To illustrate this concept, consider a simple caterpillar tree with 4 leaves. Leaf 1 is a neighbor to leaf 2, and leaf 3 is a neighbor to leaf 4, but leaf 2 is not a neighbor to leaf 3. There are four ways to correctly reconstruct the tree, all of which begin by either joining leaves 1 and 2 or leaves 3 and 4 on the first step (see Figure 2).

For general caterpillar trees with n leaves, the first step of the neighbor-joining process will incorrectly join a pair of non-neighboring leaves unless the NJ criterion $\hat{D}_{ij}$ is minimized by either $\hat{D}_{1,2}$ or $\hat{D}_{n-1,n}$. By symmetry, $\hat{D}_{1,2}$ and $\hat{D}_{n-1,n}$ are identically distributed random variables, and so we focus our attention on the behavior of $\hat{D}_{1,2}$. For the NJ criterion to be minimized for the neighboring sequences S_1 and S_2, we must have $\hat{D}_{1,2} \leq \hat{D}_{i,j}$ for *all* pairs of non-neighboring sequences S_i and S_j. Thus, if we consider a single pair of non-neighboring sequences S_{g_n} and S_{g_n+1}, the probability that the NJ criterion is not minimized by $\hat{D}_{1,2}$ will clearly be at least as large as the probability that $\hat{D}_{1,2} > \hat{D}_{g_n,g_n+1}$. Choosing g_n to be sufficiently large that the estimates $\hat{p}_{1k}$ and $\hat{p}_{g_n,k}$ are nearly independent for all k (that is, $g_n = n^\gamma$ for any $\gamma \in \left(0, \frac{1}{2}\right)$), we analyze the asymptotic properties of the random variable $D_n = \left(\hat{D}_{g_n,g_n+1} - \hat{D}_{1,2}\right)$.

2.5 Modelling details

To further simplify the analysis, we reduce the complexity of observed sequences to consist of binary strings, so that the relationship between the expected Hamming distance and the Jukes-Cantor distance now becomes $d_{ij} = -\frac{1}{2}\log(1-2p_{ij})$, $p_{ij} \in [0, \frac{1}{2})$ (this is known as the *Cavender-Farris* model [22, 23]). For this additive model, the true distance between taxa i and j, $i < j$, is given by $(j-i)d_e$, and thus the expected Hamming distance will be equal to $\frac{1}{2}\left(1-(1-2p_e)^{j-i}\right)$ where p_e is the true probability of observing a mutation on a single edge.

In this setting, it is clear that for distant sequences (where j is much larger than i), the expected Hamming distance p_{ij} will approach the critical value of $\frac{1}{2}$ exponentially fast as $j-i$ increases. On the other hand, for observed sequences that are only polynomially long, a calculation with the Binomial distribution shows that the variance of $\hat{p}_{ij}$ approaches 0 at most only polynomially fast. A Normal approximation to the Binomial then shows that, for pairs of taxa (i, j) separated enough so that $\frac{j-i}{\log n} \to \infty$, the probability $P\{\hat{p}_{ij} > \frac{1}{2}\}$ converges to $\frac{1}{2}$. That is, with probability approaching $\frac{1}{2}$ for each distantly separated pair of sequences, the standard distance estimate will be undefined.

Thus, we can easily conclude that if one does not allow for the correction of undefined distance estimates, NJ will fail to reconstruct the caterpillar tree for large values of n. The question of whether correction of undefined distances can enable NJ to succeed is more subtle. To address this, we allow for the correction of such values by assigning the maximum observable value for the sequence length L_n to any undefined distance. Assume, for definiteness, that

L_n is an even integer. We define the value of corrected distances to be $d^\star$, where

$$d^\star = -\frac{1}{2}\log\left(1 - 2\left(\frac{1}{2} - \frac{1}{L_n}\right)\right) = \frac{1}{2}\log\left(\frac{L_n}{2}\right). \quad (3)$$

We note that our results do not depend upon this particular choice, although we do assume that any "corrected" distance values will be at least as large as the maximum well-defined pairwise distance estimate for a given set of sequences.

3 Results

To analyze the behavior of the random variable D_n, we derive bounds for its expectation and variance. We find that, for sequences of polynomial length in n, the signal-to-noise ratio of D_n asympotically approaches 0: that is, the standard deviation of the distribution of D_n grows more quickly that its mean. This implies that, in the limit, observed values of D_n are equally likely to be positive or negative, and in the latter case the algorithm would incorrectly join the pair of non-neighboring leaves. Detailed proofs of the expectation and variance bounds are provided in the Appendix.

3.1 Derivation of an upper bound for the expectation of D_n

Let the notation $\hat{d}_{i\cdot} = \sum_{k=1}^{n} \hat{d}_{i,k}$. For any value g_n, the expectation of $D_n = \hat{D}_{g_n,g_n+1} - \hat{D}_{1,2}$ is equal to

$$E\left(\hat{d}_{g_n,g_n+1} - \frac{1}{n-2}\left(\hat{d}_{g_n\cdot} + \hat{d}_{g_n+1\cdot}\right) - \hat{d}_{12} + \frac{1}{n-2}\left(\hat{d}_{1\cdot} + \hat{d}_{2\cdot}\right)\right)$$
$$= E\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}\right) + \frac{1}{n-2}\left(E\left(\hat{d}_{1\cdot} - \hat{d}_{g_n\cdot}\right) + E\left(\hat{d}_{2\cdot} - \hat{d}_{g_n+1\cdot}\right)\right), \qquad (4)$$

and, because we have defined $d_{1,2} = d_e = d_{g_n,g_n+1}$, the above expression reduces to

$$E\left(\hat{D}_{g_n,g_n+1} - \hat{D}_{1,2}\right) = \frac{1}{n-2}\left(E\left(\hat{d}_{1\cdot} - \hat{d}_{g_n\cdot}\right) + E\left(\hat{d}_{2\cdot} - \hat{d}_{g_n+1\cdot}\right)\right). \qquad (5)$$

Let $\Delta_k = \left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right)$ and $\Delta'_k = \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)$. If sufficiently long sequences were available to guarantee accurate estimation of the distances $d_{1,k}$ and $d_{g_n,k}$ for all values of k, then the expectation of the sums of differences Δ_k would approach

$$\sum_{k=1}^{n}(d_{1,k} - d_{g_n,k}) = (n - g_n)(g_n - 1)d_e,$$

where d_e denotes the length of a single edge on the caterpillar tree. However, for sequences of polynomial length, we find that the expectation of the sums of differences Δ_k and Δ'_k will be significantly smaller.

To bound this expectation, we divide the sums into three segments. The first includes those terms for which $k \leq g_n + 1$, the second includes terms for which k is greater than (but relatively close to) $g_n + 1$, and the remaining segment includes the terms for which k is significantly larger than g_n. We denote the length of the middle segment by b_n and let $b_n = n^\beta$ for some $\beta \in \left(0, \frac{1}{2}\right)$. With this approach, we derive the following bound:

Theorem 1 *For n sequences of length $L = n^s$ for any fixed s,*

$$E(\hat{D}_{g_n,g_n+1} - \hat{D}_{1,2}) \leq \frac{\ln(n)}{n-2}\left(sn^{\beta} + o(n^{-1})\right)$$

for $g_n > 2$ and $\beta \in (0, \frac{1}{2})$.

3.2 Derivation of a lower bound for the variance of D_n

By definition,

$$\mathrm{Var}(D_n) = \mathrm{Var}\left[\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{1,2}\right) + \frac{1}{n-2}\left(\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right) + \left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)\right)\right]. \tag{6}$$

To analyze this expression, we first show that, because $d_{1,2} = d_e = d_{g_n,g_n+1}$ can be estimated with great precision by sequences of polynomial length, the variance of $\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{1,2}\right)$ converges to 0 for sequences of length $L_n = n^s$. It follows from this result that the covariance terms involving $\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{1,2}\right)$ also converge to 0 for polynomial length sequences (see Lemma 14). Therefore, we focus our attention on the expression

$$\begin{aligned}
&\mathrm{Var}\left(\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right) + \left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)\right) \\
=&\mathrm{Var}\left(\sum_{k=1}^{n}\Delta_k\right) + \mathrm{Var}\left(\sum_{k=1}^{n}\Delta'_k\right) + 2\mathrm{Cov}\left(\sum_{k=1}^{n}\Delta_k, \sum_{k=1}^{n}\Delta'_k\right)
\end{aligned} \tag{7}$$

where

$$\mathrm{Var}\left(\sum_{k=1}^{n}\Delta_k\right) = \sum_{k=1}^{n}\mathrm{Var}\left(\Delta_k\right) + 2\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\mathrm{Cov}\left(\Delta_k, \Delta_l\right). \tag{8}$$

When k is significantly larger than g_n, both $p_{1,k}$ and $p_{g_n,k}$ are so close to $\frac{1}{2}$ that the true proportion of differences between sequences 1 and k cannot be

accurately estimated by $L_n = n^s$ positions (in other words, we have $\frac{1}{2} - \frac{1}{n^s} < p_{g_n,k} < p_{1,k} < \frac{1}{2}$). In this case, where $g_n = n^\gamma$ for any $\gamma \in \left(0, \frac{1}{2}\right)$, the estimates $\hat{d}_{1,k}$ and $\hat{d}_{g_n,k}$ are likely to be "corrected" to the value $d^\star$ with probability close to $\frac{1}{2}$, and, because g_n is large, the estimates $\hat{d}_{1,k}$ and $\hat{d}_{g_n,k}$ are nearly independent. Because of this inability to precisely estimate the parameters $\hat{p}_{1,k}$ and $\hat{p}_{g_n,k}$ when k is large, the variability of the difference terms $\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right)$ is considerable, and Theorem 2 provides a lower bound for the variance of each term:

Theorem 2 *For $k > g_n$ and for n sequences of length $L = n^s$ for any fixed s, if $g_n = n^\gamma$ for any $\gamma \in \left(0, \frac{1}{2}\right)$, then*

$$Var(\hat{d}_{1,k} - \hat{d}_{g_n,k}) \geq \left(\frac{1}{4} - \frac{\delta_{g_n,k}}{2} - o(n^{-1})\right)\left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n}\right)^2 (\ln(n))^2$$

with $\delta_{g_n,k} = P\left(\hat{p}_{g_n,k} < \frac{1}{2}\right) - \frac{1}{2}$.

With this result, for $k > g_n + b_n$ we have

$$\mathrm{Var}\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) \geq c_{\beta,s,n} (\ln(n))^2$$

where $c_{\beta,s,n} = \left(\frac{1}{4} - \frac{\delta_{g_n,k}}{2} - o(n^{-1})\right)\left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n}\right)^2 \to \frac{1}{4}\left(\frac{s}{4} - \frac{1}{2}\right)^2$ as $n \to \infty$, and therefore

$$\sum_{k=1}^{n} \mathrm{Var}\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) > \left(n - \left(n^\gamma + n^\beta\right)\right) c_{\beta,s,n} (\ln(n))^2 \text{ and}$$

$$\sum_{k=1}^{n} \mathrm{Var}\left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right) > \left(n - \left(n^\gamma + n^\beta\right)\right) c_{\beta,s,n} (\ln(n))^2 . \tag{9}$$

Now we consider the covariance terms in Equations (7) and (8). If we assume that the contribution of the covariance terms is positive, then we may eas-

ily bound the standard deviation of D_n using Inequality (9). However, given that the covariance terms may be negative, thereby reducing the variance, we derive bounds for these terms which account for their largest possible impact. We find upper and lower bounds for the covariance of each pair of distances in Lemma 9, establishing that for all sequences S_i, S_j, S_k, and S_l, $\mathrm{Cov}\left(\hat{d}_{i,j}, \hat{d}_{k,l}\right) \geq 0$. We recognize, then, that each covariance term is bounded below:

$$\begin{aligned}
\mathrm{Cov}\left(\Delta_k, \Delta_l\right) &\geq -\left(\mathrm{Cov}\left(\hat{d}_{1,k}, \hat{d}_{g_n,l}\right) + \mathrm{Cov}\left(\hat{d}_{g_n,k}, \hat{d}_{1,l}\right)\right) \\
\mathrm{Cov}\left(\Delta'_k, \Delta'_l\right) &\geq -\left(\mathrm{Cov}\left(\hat{d}_{2,k}, \hat{d}_{g_n+1,l}\right) + \mathrm{Cov}\left(\hat{d}_{g_n+1,k}, \hat{d}_{2,l}\right)\right) \\
\text{and } \mathrm{Cov}\left(\Delta_k, \Delta'_l\right) &\geq -\left(\mathrm{Cov}\left(\hat{d}_{1,k}, \hat{d}_{g_n+1,l}\right) + \mathrm{Cov}\left(\hat{d}_{g_n,k}, \hat{d}_{2,l}\right)\right) \qquad (10)
\end{aligned}$$

We employ this fact to derive bounds for the covariance terms in Theorem 3 and Corollary 4.

Theorem 3

$$\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{Cov}\left(\Delta_k, \Delta'_l\right) \geq -\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2 \left[g_n^2 + o\left(n^{-1}\right)\right]$$

Corollary 4

$$\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} \mathrm{Cov}\left(\Delta_k, \Delta_l\right) \geq -\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2 \left[\frac{g_n^2}{2} + o\left(n^{-1}\right)\right]$$

$$\textit{and } \sum_{k=1}^{n-1}\sum_{l=k+1}^{n} \mathrm{Cov}\left(\Delta'_k, \Delta'_l\right) \geq -\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2 \left[\frac{g_n^2}{2} + o\left(n^{-1}\right)\right]$$

These inequalities demonstrate that, even in the most extreme case, the contribution of the covariance terms is negligible relative to the overall variance. Aggregating the preceding results, we derive the following lower bound:

Theorem 5 *For $\beta \in (0,1)$, $\gamma \in \left(0, \frac{1}{2}\right)$ and sequences of length $L_n = n^s$,*

$$\mathrm{Var}(D_n) \geq \left(\frac{\ln(n)}{n-2}\right)^2 \left[2\left(n - \left(n^\gamma + n^\beta\right)\right) c_{\beta,s,n} - s^2 n^{2\gamma} - o\left(n^{-1}\right)\right]$$

3.3 Relationship between expectation and variance results

Reviewing the inequalities derived Sections 3.1 and 3.2, we see that, for any $\beta \in (0,1)$ and for any $\gamma \in \left(0, \frac{1}{2}\right)$,

$$E(D_n) \leq \frac{\ln(n)}{n-2}\left(sn^\beta + o\left(n^{-1}\right)\right)$$

for sequences of polynomial length n^s, while

$$\mathrm{Var}(D_n) \geq \left(\frac{\ln(n)}{n-2}\right)^2 \left[2\left(n - \left(n^\gamma + n^\beta\right)\right) c_{\beta,s,n} - s^2 n^{2\gamma} - o\left(n^{-1}\right)\right].$$

Ignoring constants and those terms which converge to 0, we take the ratio of the expectation inequality and the standard deviation inequality:

$$\begin{aligned} \frac{E(D_n)}{\mathrm{SD}(D_n)} &\leq \frac{n^\beta}{\sqrt{n - (n^\gamma + n^\beta + n^{2\gamma})}} \\ &= \frac{1}{n^{\frac{1}{2}-\beta}\sqrt{1 - (n^{-(1-\gamma)} + n^{-(1-\beta)} + n^{-(1-2\gamma)})}}. \end{aligned} \tag{11}$$

This ratio will approach 0 in the limit as $n \to \infty$ for any $\beta,\ \gamma \in \left(0, \frac{1}{2}\right)$, indicating that, for a wide range of possible values of g_n, the variability of the difference $\left(\hat{D}_{g_n,g_n+1} - \hat{D}_{1,2}\right)$ is increasing much more rapidly than its expected value. Furthermore, if we assume that the distribution of D_n is reasonably

well-behaved, then

$$\lim_{n\to\infty} P(D_n < 0) = \lim_{n\to\infty} P\left(\frac{D_n - E(D_n)}{\sigma_{D_n}} < \frac{-E(D_n)}{\sigma_{D_n}}\right) = \frac{1}{2}. \qquad (12)$$

4 Discussion

By analyzing the random variable $D_n = \left(\hat{D}_{g_n,g_n+1} - \hat{D}_{12}\right)$, we have shown that the NJ criterion is likely to be minimized by a pair of non-neighboring leaves when polynomial length sequences are considered. Our results demonstrate the vulnerability of the method to the impact of large numbers of imprecise distance estimates, as reflected in the asymptotic behavior of the signal-to-noise ratio of D_n. It is therefore apparent that polynomial length sequences will be insufficient to guarantee phylogenetic accuracy for at least one class of trees, and that Atteson's exponential bound cannot be improved in general.

Our theoretical result should not necessarily be perceived as an indictment of the value of NJ as a practical phylogeny reconstruction method. The caterpillar tree considered here represents an extreme case, rarely (if ever) encountered in a realistic biological setting. As demonstrated by numerous simulation studies that have considered more typical trees (including [24, 25, 26, 27]), NJ does in fact perform quite well with reasonably short sequence lengths.

The difficulties in phylogeny reconstruction demonstrated by the present analysis are not simply a failing of the NJ algorithm but rather arise from an interaction between the NJ algorithm and the "fast convergence" criterion. The criterion of fast convergence requires consideration of trees whose number of taxa tends to infinity while maintaining a fixed positive lower bound on the

edge lengths in the tree. This in turn forces the existence of ever more remotely separated pairs of taxa. More natural alternative asymptotic formulations of the process of biologists collecting data on more and more species might assume a bounded time since the root of the tree, in which case the minimum branch length of the tree would approach 0 as $n \to \infty$. Characterizing the performance of algorithms including NJ in such a framework is an area for further research.

More practically, the results presented in this analysis demonstrate the problems that can arise when sequence lengths are insufficiently long to estimate large distances with precision. For any phylogeny involving very distantly related taxa, it is to be expected that a significant number of pairwise distance estimates will be inaccurate or undefined, and our analysis demonstrates that NJ is highly susceptible to these errors. The insights drawn from this study are therefore not restricted to the artificial special case of the caterpillar topology, but rather can be extended to a much larger class of phylogeny reconstruction problems.

A Proofs for Expectation and Variance Bounds

The following Lemma is employed in the Proof of Theorem 1.

Lemma 6

$$\sum_{k=g_n+b_n+1}^{n} E\left(\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) + \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)\right) < \frac{\ln(n)}{p_e} s n^s (1-2p_e)^{b_n}.$$

Proof of Lemma 6 For integers $i \in [0, L_n]$, let $d(i)$ denote the distance function

$$d(i) = \begin{cases} -\frac{1}{2}\ln(1-\frac{2i}{L_n}), \;\; 0 \le i < \frac{L_n}{2}, \\ \\ d^\star = \frac{1}{2}\ln\left(\frac{L_n}{2}\right), \; i \ge \frac{L_n}{2} \end{cases}$$

Then for any $k > g_n$, we have

$$\begin{aligned}
& E(\hat{d}_{1,k} - \hat{d}_{g_n,k}) \\
&= \sum_{i=0}^{L_n}\sum_{j=0}^{L_n} P\left(\hat{p}_{1,k} = \frac{i}{L_n}, \hat{p}_{g_n,k} = \frac{j}{L_n}\right)\Big[(d(i)-d(j))\left\{i,j < \frac{L_n}{2}\right\} \\
&\quad + (d^\star - d(j))\left\{i \ge \frac{L_n}{2}, j < \frac{L_n}{2}\right\} + (d(i)-d^\star)\left\{i < \frac{L_n}{2}, j \ge \frac{L_n}{2}\right\}\Big] \\
&= \sum_{i=0}^{L_n-1}\sum_{j=i+1}^{L_n}\left(P\left(\hat{p}_{1,k} = \frac{j}{L_n}, \hat{p}_{g_n,k} = \frac{i}{L_n}\right) - P\left(\hat{p}_{1,k} = \frac{i}{L_n}, \hat{p}_{g_n,k} = \frac{j}{L_n}\right)\right) \times \\
&\quad \Big[\left(-\frac{1}{2}\ln\left(\frac{L_n-2j}{L_n-2i}\right)\right)\left\{i,j < \frac{L_n}{2}\right\} + (d^\star - d(i))\left\{i < \frac{L_n}{2}, j \ge \frac{L_n}{2}\right\}\Big] \\
&\le d^\star \sum_{i=0}^{L_n-1}\sum_{j=i+1}^{L_n}\left(P\left(\hat{p}_{1,k} = \frac{j}{L_n}, \hat{p}_{g_n,k} = \frac{i}{L_n}\right) - P\left(\hat{p}_{1,k} = \frac{i}{L_n}, \hat{p}_{g_n,k} = \frac{j}{L_n}\right)\right).
\end{aligned} \tag{A.1}$$

For a single position r, let $Y_{1k,r} = 1$ if $S_{1,r} \neq S_{k,r}$, 0 otherwise, and let $Y_{g_nk,r} = 1$ if $S_{g_n,r} \neq S_{k,r}$, 0 otherwise. It follows that $Y_{1k,r}$ and $Y_{g_nk,r}$ are dependent Bernoulli($p_{1,k}$) and Bernoulli($p_{g_n,k}$) random variables. For $1 \le l \le L_n$, define $X_{1k}^{(l)} = \sum_{r=1}^{l} Y_{1k,r}$, and $X_{g_nk}^{(l)} = \sum_{r=1}^{l} Y_{g_nk,r}$. We note that

$$\begin{aligned}
& \sum_{i=0}^{L_n-1}\sum_{j=i+1}^{L_n}\left(P\left(\hat{p}_{1,k} = \frac{j}{L_n}, \hat{p}_{g_n,k} = \frac{i}{L_n}\right) - P\left(\hat{p}_{1,k} = \frac{i}{L_n}, \hat{p}_{g_n,k} = \frac{j}{L_n}\right)\right) \\
&= \sum_{i=0}^{L_n-1}\sum_{j=i+1}^{L_n}\left(P(X_{1k}^{(L_n)} = j, X_{g_nk}^{(L_n)} = i) - P(X_{1k}^{(L_n)} = i, X_{g_nk}^{(L_n)} = j)\right) \\
&< \sum_{i=0}^{L_n}\sum_{j=0}^{L_n}\left|P(X_{1k}^{(L_n)} = j, X_{g_nk}^{(L_n)} = i) - P(X_{1k}^{(L_n)} = i, X_{g_nk}^{(L_n)} = j)\right|,
\end{aligned} \tag{A.2}$$

and, by properties of the total variation distance (see Section A.1),

$$\sum_{i=0}^{L_n}\sum_{j=0}^{L_n}\left|P(X_{1k}^{(L_n)}=j,X_{g_nk}^{(L_n)}=i)-P(X_{1k}^{(L_n)}=i,X_{g_nk}^{(L_n)}=j)\right|$$
$$\leq (L_n)\sum_{i=0}^{1}\sum_{j=0}^{1}\left|P(Y_{1k,r}=j,Y_{g_nk,r}=i)-P(Y_{1k,r}=i,X_{g_nk,r}=j)\right|. \quad \text{(A.3)}$$

Let $p_{1k,g_nk}(i,j)=P(Y_{1k,r}=i,Y_{g_nk,r}=j)$, and let $q_{1k,g_nk}(i,j)=P(Y_{1k,r}=j,Y_{g_nk,r}=i)$ for $i,j\in 0,1$. Assume, without loss of generality, that sequence $S_{1,r}=0$ for all positions r. We bound the distance $\sum_{i,j}|p_{1k,g_nk}(i,j)-q_{1k,g_nk}(i,j)|$. Because $p_{1k,g_nk}(i,i)=q_{1k,g_nk}(i,i)$, we need only evaluate one of the two cases where $i\neq j$ (by symmetry, $|p_{1k,g_nk}(1,0)-q_{1k,g_nk}(1,0)|=|p_{1k,g_nk}(0,1)-q_{1k,g_nk}(0,1)|$). The probability $p_{1k,g_nk}(1,0)=P(Y_{1k,r}=1,Y_{g_nk,r}=0)$ is given by

$$p_{1k,g_nk}(1,0)=P\left(S_{1,r}=0,S_{g_n,r}=1,S_{k,r}=1\right)=p_{1,g_n}(1-p_{g_n,k}), \quad \text{(A.4)}$$

and the probability $q_{1k,g_nk}(1,0)=P(Y_{1k,r}=0,Y_{g_nk,r}=1)$ is given by

$$q_{1k,g_nk}(1,0)=P\left(S_{1,r}=0,S_{g_n,r}=1,S_{k,r}=0\right)=p_{1,g_n}p_{g_n,k}. \quad \text{(A.5)}$$

It follows that

$$|p_{1k,g_nk}(1,0)-q_{1k,g_nk}(1,0)|=p_{1,g_n}(1-2p_{g_n,k})<\frac{1}{2}(1-2p_e)^{k-g_n}, \quad \text{(A.6)}$$

and with this bound we have, for each k,

$$\begin{aligned}
&E(\hat{d}_{1,k} - \hat{d}_{g_n,k}) \\
&\leq d^\star \sum_{i=0}^{L_n-1} \sum_{j=i+1}^{L_n} \left(P\left(\hat{p}_{1,k} = \frac{j}{L_n}, \hat{p}_{g_n,k} = \frac{i}{L_n}\right) - P\left(\hat{p}_{1,k} = \frac{i}{L_n}, \hat{p}_{g_n,k} = \frac{j}{L_n}\right)\right) \\
&< L_n(1-2p_e)^{k-g_n} \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right). \qquad \text{(A.7)}
\end{aligned}$$

Summing over all of the terms, we find

$$\begin{aligned}
E\left(\sum_{k=g_n+b_n+1}^{n} \hat{d}_{1,k} - \hat{d}_{g_n,k}\right) &= \frac{L_n}{2}\ln\left(\frac{L_n}{2}\right) \sum_{k=g_n+b_n+1}^{n} (1-2p_e)^{k-g_n} \\
&= \frac{L_n}{2}\ln\left(\frac{L_n}{2}\right)(1-2p_e)^{b_n+1} \sum_{k=0}^{n-(g_n+b_n+1)} (1-2p_e)^k \\
&= \frac{L_n}{2}\ln\left(\frac{L_n}{2}\right)(1-2p_e)^{b_n+1}\left(\frac{1-(1-2p_e)^{n-(g_n+b_n)}}{1-(1-2p_e)}\right) \\
&< \frac{\ln(n)}{2p_e} s n^s (1-2p_e)^{b_n} \qquad \text{(A.8)}
\end{aligned}$$

for sequences of length $L_n = n^s$. To bound the expectation of the sum of differences $\left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)$, we note that for any $k > g_n + 1$

$$E\left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right) = E\left(\hat{d}_{1,k-1} - \hat{d}_{g_n,k-1}\right), \qquad \text{(A.9)}$$

and so we also find

$$E\left(\sum_{k=g_n+b_n+1}^{n} \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)\right) < \frac{\ln(n)}{2p_e} s n^s (1-2p_e)^{b_n} \qquad \text{(A.10)}$$

as above. □

Proof of Theorem 1 By definition, for any $g_n > 2$, the expectation of $\hat{D}_{g_n,g_n+1} - \hat{D}_{12}$ is equal to

$$\begin{aligned} & E\left(\hat{d}_{g_n,g_n+1} - \frac{1}{n-2}\left(\hat{d}_{g_n\cdot} + \hat{d}_{g_n+1\cdot}\right) - \hat{d}_{12} + \frac{1}{n-2}\left(\hat{d}_{1\cdot} + \hat{d}_{2\cdot}\right)\right) \\ = & \frac{1}{n-2}\left(E\left(\hat{d}_{1\cdot} - \hat{d}_{g_n\cdot}\right) + E\left(\hat{d}_{2\cdot} - \hat{d}_{g_n+1\cdot}\right)\right) \end{aligned} \tag{A.11}$$

by the equality of d_{12} and d_{g_n,g_n+1}.

To derive the result, we divide the terms into three regions. The first region is defined for $k \in [1, g_n + 1]$, the second region is defined by $k \in [g_n + 2, g_n + b_n]$ where $b_n = n^\beta$ for any $\beta \in \left(0, \frac{1}{2}\right)$, and the final region includes those terms for which $k > g_n + b_n$.

For the first region we find that

$$\sum_{k=1}^{g_n+1} E\left(\left(\hat{d}_{1,k} - \hat{d}_{g_n k}\right) + \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)\right) = 0, \tag{A.12}$$

an intuitive result that is easily verified by direct calculations.

For the remaining terms, we first consider those terms for which k is relatively close to g_n, with $k \in [g_n + 2, g_n + b_n]$. In this region, the mutation probabilities $p_{1k} = \frac{1}{2}\left(1 - (1-2p_e)^{k-1}\right) > \frac{1}{2}\left(1 - (1-2p_e)^{g_n}\right)$, and because $g_n = n^\gamma$ it is clear that, for sequences of polynomial length, many estimates $\hat{p}_{1,k}$ and $\hat{p}_{2k}$ will be greater than or equal to $\frac{1}{2}$. However, some of the mutation probabilities $p_{g_n,k}$ will not be large in this region, and we can expect, for k close to g_n, that the distance estimates $\hat{d}_{g_n,k}$ and $\hat{d}_{g_n+1,k}$ will be well-defined and, therefore, less than the maximum value $d^\star$. To account for this behavior in our analysis, we bound the expectation by assuming that all estimates $\hat{d}_{1,k}$ and $\hat{d}_{2,k}$ are

assigned the maximum value $d^\star = \frac{1}{2}\log\left(\frac{L_n}{2}\right)$, while all estimates $\hat{d}_{g_n,k}$ and $\hat{d}_{g_n+1,k}$ are well-defined. It follows that, for sequences of length $L_n = n^s$ for any $s > 1$,

$$\begin{aligned}
&\sum_{k=g_n+2}^{g_n+b_n} E\left(\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) + \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)\right) \\
&< \sum_{k=g_n+2}^{g_n+b_n} \left(\left(d^\star - E(\hat{d}_{g_n,k})\right) + \left(d^\star - E(\hat{d}_{g_n+1,k})\right)\right) < 2b_n d^\star < b_n s\ln(n). \quad \text{(A.13)}
\end{aligned}$$

In the region for which $k > g_n + b_n$, the distances d_{1k} and $d_{g_n,k}$ are both sufficiently large that, for sequences of polynomial length, the distributions of $\hat{p}_{1,k}$ and $\hat{p}_{g_n,k}$ are nearly identical. We find that the expectation of the differences $\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right)$ and $\left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)$ is negligible in this region, decreasing to 0 with n as established in Lemma 6. Aggregating these results, an overall upper bound is given by

$$E(\hat{D}_{g_n,g_n+1} - \hat{D}_{12}) \leq \frac{\ln(n)}{n-2}\left(sb_n + \frac{sn^s(1-2p_e)^{b_n}}{p_e}\right), \quad \text{(A.14)}$$

and, for $b_n = n^\beta$ for any $\beta \in \left(0, \frac{1}{2}\right)$ and $p_e > 0$, $n^s(1-2p_e)^{b_n} = o(n^{-1})$. $\square$

Lemmas 7 and 8 are employed in the Proof of Theorem 2.

Lemma 7 *For any distance estimate $\hat{d}_{ij}$ with $\hat{p}_{ij} < \frac{1}{2}$,*

$$E\left(\hat{d}_{ij}\right) < \left(\frac{s}{4} + \frac{1}{2} + \frac{s}{n}\right)\ln(n) - \frac{1}{2}\ln(2).$$

Proof of Lemma 7 To establish this result, we first show that $P\left(\hat{p}_{ij} \geq \frac{1}{2} - \frac{1}{n\sqrt{L_n}}\right) < \frac{2}{n}$. Note that, for any well-defined estimate $\hat{p}_{ij}$,

$$P\left(\hat{p}_{ij} \geq \frac{1}{2} - \frac{1}{n\sqrt{L_n}}\right) = P\left(\hat{p}_{ij} \in \left[\frac{1}{2} - \frac{1}{n\sqrt{L_n}}, \frac{1}{2}\right)\right). \tag{A.15}$$

Let $\hat{p}_{ij} = \frac{1}{L_n} X_{ij}$, where $X_{ij} = \sum_{r=1}^{L_n} (S_{i,r} \neq S_{j,r}) \sim B\left(L_n, p_{ij}\right)$. Assume, without loss of generality, that L_n is an even integer. Then

$$P\left(\frac{L_n}{2} - \frac{\sqrt{L_n}}{n} \leq X_{ij} < \frac{L_n}{2}\right) < P\left(\frac{L_n}{2} - \left\lceil \frac{\sqrt{L_n}}{n} \right\rceil \leq X_{ij} < \frac{L_n}{2}\right), \tag{A.16}$$

where $\left\lceil \frac{\sqrt{L_n}}{n} \right\rceil$ denotes the smallest integer greater than $\frac{\sqrt{L_n}}{n}$. Let X' be a binomial random variable with size L_n and probability $p' = \frac{1}{2} - \frac{1}{L_n}\left\lceil \frac{\sqrt{L_n}}{n} \right\rceil$. Of all binomial random variables with size L_n and probability $p \in \left[\frac{1}{2} - \frac{1}{L_n}\left\lceil \frac{\sqrt{L_n}}{n} \right\rceil, \frac{1}{2}\right)$, the distribution of X' has the smallest variance. Thus, the maximum value of the probability mass function of X', which is achieved at its expected value, is greater than the maximum value achieved by any other binomial random variable of size L_n on the interval $\left[\frac{L_n}{2} - \left\lceil \frac{\sqrt{L_n}}{n} \right\rceil, \frac{L_n}{2}\right)$. It follows that

$$P\left(\frac{L_n}{2} - \left\lceil \frac{\sqrt{L_n}}{n} \right\rceil \leq X_{ij} < \frac{L_n}{2}\right) < P\left(X' = \frac{L_n}{2} - \left\lceil \frac{\sqrt{L_n}}{n} \right\rceil\right)\left(\frac{\sqrt{L_n}}{n}\right). \tag{A.17}$$

By Stirling's approximation, $L! \sim \sqrt{2\pi} L^{\left(L+\frac{1}{2}\right)} e^{-L}$. Then for any $B(L,p)$ random variable X for which Lp is an integer,

$$\begin{aligned} P\left(X = Lp\right) &= \frac{L!}{(Lp)!(L(1-p))!} p^{Lp}(1-p)^{L(1-p)} \\ &\sim \frac{L^{L+\frac{1}{2}}}{\sqrt{2\pi}(Lp)^{Lp+\frac{1}{2}}(L(1-p))^{L(1-p)+\frac{1}{2}}} \left(p^{Lp}(1-p)^{L(1-p)}\right) \\ &= \left(2\pi Lp(1-p)\right)^{-\frac{1}{2}}. \end{aligned} \tag{A.18}$$

With this result, we have

$$P\left(X' = \frac{L_n}{2} - \left\lceil \frac{\sqrt{L_n}}{n} \right\rceil\right) \sim \left(2\pi L_n \left(\frac{1}{2} - \frac{1}{L_n}\left\lceil \frac{\sqrt{L_n}}{n} \right\rceil\right)\left(\frac{1}{2} + \frac{1}{L_n}\left\lceil \frac{\sqrt{L_n}}{n} \right\rceil\right)\right)^{-\frac{1}{2}}$$
$$\leq \left(2\pi L_n \left(\frac{1}{2} - \frac{1}{L_n}\left\lceil \frac{\sqrt{L_n}}{n} \right\rceil\right)^2\right)^{-\frac{1}{2}} < \frac{2}{\sqrt{L_n}} \qquad \text{(A.19)}$$

for all $s \geq 1$ and $n \geq 4$, and it follows directly that

$$P\left(\hat{p}_{ij} \geq \frac{1}{2} - \frac{1}{\sqrt{L_n} n}\right) < P\left(X' = \frac{L_n}{2} - \left\lceil \frac{\sqrt{L_n}}{n} \right\rceil\right)\left(\frac{\sqrt{L_n}}{n}\right) < \frac{2}{n}. \qquad \text{(A.20)}$$

By definition, $\hat{d}_{ij} = -\frac{1}{2}\ln(1 - 2\hat{p}_{ij})$ for all $\hat{p}_{ij} < \frac{1}{2}$. Then, for even integers L_n, the maximum possible value for $\hat{d}_{ij} = d^\star = -\frac{1}{2}\ln\left(1 - 2\left(\frac{1}{2} - \frac{1}{L_n}\right)\right) = \frac{1}{2}\ln\left(\frac{L_n}{2}\right)$, and so we may conservatively bound the expectation of $\hat{d}_{ij}$ with the expression

$$E(\hat{d}_{ij}) \leq -\frac{1}{2}\ln\left(1 - 2\left(\frac{1}{2} - \frac{1}{n\sqrt{L_n}}\right)\right)\left(1 - \frac{2}{n}\right) + \frac{1}{2}\ln\left(\frac{L_n}{2}\right)\left(\frac{2}{n}\right)$$
$$= \left(\frac{1}{2} - \frac{1}{n}\right)\ln\left(\frac{n\sqrt{L_n}}{2}\right) + \left(\frac{1}{n}\right)\ln\left(\frac{L_n}{2}\right). \qquad \text{(A.21)}$$

The final result follows from simplifying the expression for sequences of length $L_n = n^s$. $\square$

Lemma 8 *If $g_n = n^\gamma$ for any $\gamma \in \left(0, \frac{1}{2}\right)$, then*

$$P\left(\hat{p}_{1,k} < \frac{1}{2}, \hat{p}_{g_n,k} \geq \frac{1}{2}\right) > \frac{1}{4} - \frac{1}{2}\left(\delta_{g_n,k} + o(n^{-1})\right),$$

where $\delta_{g_n,k} = P\left(\hat{p}_{g_n,k} < \frac{1}{2}\right) - \frac{1}{2}$.

Proof of Lemma 8 For any $k > g_n$, the probability $P\left(\hat{p}_{1,k} < \frac{1}{2}\right) = \frac{1}{2} + \delta_{1k}$, with $\delta_{1k} \to 0$ as $k \to \infty$, and, similarly, $P\left(\hat{p}_{g_n,k} < \frac{1}{2}\right) = \frac{1}{2} + \delta_{g_n k}$. Letting $X_{1k} = L_n \hat{p}_{1,k}$ and $X_{g_n k} = L_n \hat{p}_{g_n,k}$, we have

$$P\left(\hat{p}_{1,k} < \frac{1}{2}, \hat{p}_{g_n,k} \geq \frac{1}{2}\right) = P\left(X_{1k} < \frac{L_n}{2}, X_{g_n k} \geq \frac{L_n}{2}\right).$$

Because X_{1k} and $X_{g_n k}$ will be nearly independent for large k, we derive a bound by analyzing the difference between the joint probability and the product of the marginal probabilities,

$$\begin{aligned}
&\left|P\left(X_{1k} < \frac{L_n}{2}, X_{g_n k} \geq \frac{L_n}{2}\right) - P\left(X_{1k} < \frac{L_n}{2}\right) P\left(X_{g_n k} \geq \frac{L_n}{2}\right)\right| \\
&= \left|P\left(X_{1k} < \frac{L_n}{2}\right)\left[P\left(X_{g_n k} \geq \frac{L_n}{2} \,\middle|\, X_{1k} < \frac{L_n}{2}\right) - P\left(X_{g_n k} \geq \frac{L_n}{2}\right)\right]\right| \\
&\leq \sum_{i=0}^{\frac{L_n}{2}-1} P(X_{1k} = i) \sum_{j=0}^{L_n} \left|P\left(X_{g_n k} = j \,|X_{1k} = i\right) - P\left(X_{g_n k} = j\right)\right|. \qquad \text{(A.22)}
\end{aligned}$$

By Lemma 11,

$$\sum_{j=0}^{L_n} \left|P\left(X_{g_n k} = j \,|X_{1k} = i\right) - P\left(X_{g_n k} = j\right)\right| \leq \frac{L_n (1-2p_e)^{g_n+1}}{1-(1-2p_e)^2}, \qquad \text{(A.23)}$$

and thus

$$\begin{aligned}
&P\left(X_{1k} < \frac{L_n}{2}, X_{g_n k} \geq \frac{L_n}{2}\right) \\
&\geq P\left(X_{1k} < \frac{L_n}{2}\right)\left[P\left(X_{g_n k} \geq \frac{L_n}{2}\right) - \frac{L_n (1-2p_e)^{g_n+1}}{1-(1-2p_e)^2}\right] \\
&= \left(\frac{1}{2} + \delta_{1k}\right)\left(\frac{1}{2} - \delta_{g_n k} - \frac{L_n (1-2p_e)^{g_n+1}}{1-(1-2p_e)^2}\right) \\
&> \frac{1}{4} - \frac{1}{2}\left(\delta_{g_n k} + \frac{L_n (1-2p_e)^{g_n+1}}{1-(1-2p_e)^2}\right), \qquad \text{(A.24)}
\end{aligned}$$

where, for sequences of length $L_n = n^s$ and $g_n = n^\gamma$ for $\gamma \in \left(0, \frac{1}{2}\right)$,

$$\frac{L_n(1-2p_e)^{g_n+1}}{1-(1-2p_e)^2} = cn^s(1-2p_e)^{n^\gamma} = o\left(n^{-1}\right) \tag{A.25}$$

for all $p_e > 0$. $\square$

Proof of Theorem 2 For any pair of observations $\hat{d}_{1,k}$ and $\hat{d}_{g_n,k}$, let the variable Y_k be defined as follows:

$$Y_k = \begin{cases} RR \text{ if } \hat{p}_{1,k} < \frac{1}{2} \text{ and } \hat{p}_{g_n,k} < \frac{1}{2} \\ RF \text{ if } \hat{p}_{1,k} < \frac{1}{2} \text{ and } \hat{p}_{g_n,k} \geq \frac{1}{2} \\ FR \text{ if } \hat{p}_{1,k} \geq \frac{1}{2} \text{ and } \hat{p}_{g_n,k} < \frac{1}{2} \\ FF \text{ if } \hat{p}_{1,k} \geq \frac{1}{2} \text{ and } \hat{p}_{g_n,k} \geq \frac{1}{2} \end{cases} \tag{A.26}$$

Conditioning on Y_k, we have

$$\begin{aligned} \mathrm{Var}(\hat{d}_{1,k} - \hat{d}_{g_n,k}) &= E(\mathrm{Var}(\hat{d}_{1,k} - \hat{d}_{g_n,k}|Y_k)) + \mathrm{Var}(E(\hat{d}_{1,k} - \hat{d}_{g_n,k}|Y_k)) \\ &\geq \mathrm{Var}(E(\hat{d}_{1,k} - \hat{d}_{g_n,k}|Y_k)). \end{aligned} \tag{A.27}$$

We recall Inequality(A.7) in the proof of Lemma 6 which states that, for sequences of length $L_n = n^s$ for any integer s, for $k > g_n$,

$$E\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) \leq sn^s(1-2p_e)^{k-g_n}\ln(n),$$

and, since $d_{1k} \geq d_{g_nk}$ for all $g_n > 1$, $E\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) \geq 0$. By definition,

$$\begin{aligned} &\mathrm{Var}(E(\hat{d}_{1,k} - \hat{d}_{g_n,k}|Y_k)) \\ &= \sum_{i \in \{FF,FR,RF,RR\}} P(Y_k = i)\left(E(\hat{d}_{1,k} - \hat{d}_{g_n,k}|Y_k = i) - E(\hat{d}_{1,k} - \hat{d}_{g_n,k})\right)^2 \\ &\geq P(Y_k = RF)\left(\left(E\left(\hat{d}_{1,k}\middle|\hat{p}_{1,k} < \frac{1}{2}\right) - d^\star\right) - E(\hat{d}_{1,k} - \hat{d}_{g_n,k})\right)^2. \end{aligned} \tag{A.28}$$

Because $E(\hat{d}_{1,k} - \hat{d}_{g_n,k}) \geq 0$ and $E(\hat{d}_{1,k}|\hat{p}_{1,k} < \frac{1}{2}) - d^\star \leq 0$, it is clear that

$$\begin{aligned} & P(Y_k = RF) \left(\left(E\left(\hat{d}_{1,k} \middle| \hat{p}_{1,k} < \frac{1}{2} \right) - d^\star \right) - E(\hat{d}_{1,k} - \hat{d}_{g_n,k}) \right)^2 \\ \geq & P(Y_k = RF) \left(E\left(\hat{d}_{1,k} \middle| \hat{p}_{1,k} < \frac{1}{2} \right) - d^\star \right)^2 . \end{aligned} \tag{A.29}$$

By Lemma 7,we have, for $d^\star = \frac{1}{2} \ln\left(\frac{L_n}{2}\right) = \frac{s}{2}\ln(n) - \frac{1}{2}\ln(2)$,

$$\left| \left(E\left(\hat{d}_{1,k} \middle| \hat{p}_{1,k} < \frac{1}{2} \right) - d^\star \right) \right| \geq \left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n} \right) \ln(n). \tag{A.30}$$

Substituting this inequality into (A.29), it follows that

$$\mathrm{Var}(E(\hat{d}_{1,k} - \hat{d}_{g_n,k}|Y_k)) \geq P(Y_k = RF) \left(\left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n} \right) \ln(n) \right)^2 . \tag{A.31}$$

Theorem 2 follows directly from this result and Lemma 8. □

Lemmas 9, 11, 12, and 13 are employed in the Proof of Theorem 3 and Corollary 4.

Lemma 9 *For any four sequences $S_a, S_b, S_c,$, and S_d with $a \neq b$ and $c \neq d$,*

$$\begin{aligned} & 0 \leq \mathrm{Cov}(\hat{d}_{ab}, \hat{d}_{cd}) \\ & \leq \left(\frac{1}{2} \ln\left(\frac{L_n}{2} \right) \right)^2 \max_{0 \leq i \leq L_n} \sum_{j=0}^{L_n} \left| P\left(\hat{p}_{cd} = \frac{j}{L_n} \middle| \hat{p}_{ab} = \frac{i}{L_n} \right) - P\left(\hat{p}_{cd} = \frac{j}{L_n} \right) \right| . \end{aligned}$$

Proof of Lemma 9 We first establish the upper bound. Define the distance function $d(i)$ as in the Proof of Lemma 6. Then

$$\begin{aligned}&\mathrm{Cov}(\hat{d}_{ab}, \hat{d}_{cd}) \\ =&\sum_{i=0}^{L_n} P\left(\hat{p}_{ab} = \frac{i}{L_n}\right) d(i) \sum_{j=0}^{L_n} d(j) \left[P\left(\hat{p}_{cd} = \frac{j}{L_n} \middle| \hat{p}_{ab} = \frac{i}{L_n}\right) - P\left(\hat{p}_{cd} = \frac{j}{L_n}\right)\right].\end{aligned} \tag{A.32}$$

Bounding $d(i)$ and $d(j)$ by $d^\star = \frac{1}{2}\ln(\frac{L_n}{2})$ for all i and j, we have

$$\begin{aligned}&\mathrm{Cov}(\hat{d}_{ab}, \hat{d}_{cd}) \\ \leq&\sum_{i=0}^{L_n} P\left(\hat{p}_{ab} = \frac{i}{L_n}\right) \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2 \sum_{j=0}^{L_n} \left|P\left(\hat{p}_{cd} = \frac{j}{L_n} \middle| \hat{p}_{ab} = \frac{i}{L_n}\right) - P\left(\hat{p}_{cd} = \frac{j}{L_n}\right)\right| \\ \leq&\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2 \max_{0\leq i\leq L_n} \sum_{j=0}^{L_n} \left|P\left(\hat{p}_{cd} = \frac{j}{L_n} \middle| \hat{p}_{ab} = \frac{i}{L_n}\right) - P\left(\hat{p}_{cd} = \frac{j}{L_n}\right)\right|.\end{aligned}$$

The lower bound follows from Proposition 10.

Proposition 10 *For any sequence length L_n and any sequences S_a, S_b, S_c, and S_d with $a \leq b \leq c \leq d$,*

(i) $\mathrm{Cov}\left(\hat{d}_{ac}, \hat{d}_{bd}\right) \geq 0.$

(ii) $\mathrm{Cov}\left(\hat{d}_{ad}, \hat{d}_{bc}\right) \geq 0.$

Proof of Proposition 10 To establish this result, we apply a classic result from Lehmann [28].

Definitions:

(1) A pair of random variables X and Y are said to be *positively quadrant*

dependent if $P(X \geq x, Y \geq y) \geq P(X \geq x)P(Y \geq x)$ for all x, y. Let $\mathscr{F}_1$ denote the family of all distributions F satisfying this property.

(2) Two real-valued functions r and s of n arguments are said to be *concordant* for the ith coordinate, if, considered as functions of the ith coordinate (with all other coordinates held fixed), they are monotone in the same direction.

We state the relevant portion of Lehmann's results:

Theorem[28]. *Let $(X_1, Y_1), \ldots (X_n, Y_n)$ be independent pairs of random variables with joint distributions $F_1, \ldots F_n$. Let r and s be functions of n variables and let*

$$X = r(X_1, \ldots, X_n),\ Y = s(Y_1, \ldots, Y_n).$$

Then $(X, Y) \in \mathscr{F}_1$ if, for each i, $F_1 \in \mathscr{F}_1$ and r, s are concordant for the ith coordinate. Furthermore, provided the expectations $E(X)$ and $E(Y)$ exist, $(X, Y) \in \mathscr{F}_1 \Rightarrow E(XY) \geq E(X)E(Y)$.

For each position i in sequences S_a and S_b, let $X_{ab,i} = I_{(S_{a,i} \neq S_{b,i})}$. Let

$$r(X_{ab,1}, \ldots, X_{ab,L_n}) = \begin{cases} -\frac{1}{2}\ln\left(1 - \frac{2\sum_{i=1}^{L_n} X_{ab,i}}{L_n}\right), & \sum_{i=1}^{L_n} X_{ab,i} < \frac{L_n}{2} - 1, \\ \frac{1}{2}\ln\left(\frac{L_n}{2}\right), & \sum_{i=1}^{L_n} X_{ab,i} \geq \frac{L_n}{2} - 1 \end{cases}$$

We first establish that $(X_{ac,i}, X_{bd,i}) \in \mathscr{F}_1$ and $(X_{ad,i}, X_{bc,i}) \in \mathscr{F}_1$. For binary random variables X and Y, we need only check that $P(X = 1, Y = 1) \geq P(X = 1)P(Y = 1)$ to establish that $(X, Y) \in \mathscr{F}_1$. In the first case, we have

$$P(X_{ac,i} = 1, X_{bd,i} = 1) = (1 - p_{ab})(p_{bc})(1 - p_{cd}) + (p_{ab})(1 - p_{bc})(p_{cd})$$

while

$$P(X_{ac,i}=1)P(X_{bd,i}=1)=p_{ac}p_{bd}.$$

Taking the difference and simplifying, we find that

$$\begin{aligned}&P(X_{ac,i}=1,X_{bd,i}=1)-P(X_{ac,i}=1)P(X_{bd,i}=1)\\ =&(1-2p_{ab})p_{bc}(1-p_{bc})(1-2p_{cd})\geq 0\end{aligned}\tag{A.33}$$

since $0\leq p_{ij}<\frac{1}{2}$ for all sequences S_i and S_j. And in the second case, we have

$$\begin{aligned}&P(X_{ad,i}=1,X_{bc,i}=1)-P(X_{ad,i}=1)P(X_{bc,i}=1)\\ =&(1-2p_{ab})p_{bc}(1-p_{bd}-p_{cd})\geq 0.\end{aligned}\tag{A.34}$$

Noting that the concordance condition is trivially satisfied, it follows from Lehmann's Theorem that

$$(r(X_{ac,1},\ldots,X_{ac,L_n}),r(X_{bd,1},\ldots,X_{bd,L_n}))\in\mathscr{F}_1 \text{ and}$$

$$(r(X_{ad,1},\ldots,X_{ad,L_n}),r(X_{bc,1},\ldots,X_{bc,L_n}))\in\mathscr{F}_1.$$

And since $\hat{d}_{ij}=r(X_{ij,1},\ldots,X_{ij,L_n})$ for all sequences S_i and S_j, we see that $(\hat{d}_{ac},\hat{d}_{bd})\in\mathscr{F}_1$ and $(\hat{d}_{ad},\hat{d}_{bc})\in\mathscr{F}_1$ to complete the proof. □

Lemma 11 (Total Variation Distance bounds) *For three sequences S_a, S_b, and S_c with $a<b<c$,*

$$(i)\ \max_{0\leq i\leq L_n}\sum_{j=0}^{L_n}\left|P\left(\hat{p}_{ac}=\frac{j}{L_n}\,\middle|\,\hat{p}_{ab}=\frac{i}{L_n}\right)-P\left(\hat{p}_{ac}=\frac{j}{L_n}\right)\right|\leq 2L_n(1-2p_e)^{c-b}$$

$$(ii)\ \max_{0\leq i\leq L_n}\sum_{j=0}^{L_n}\left|P\left(\hat{p}_{bc}=\frac{j}{L_n}\,\middle|\,\hat{p}_{ac}=\frac{i}{L_n}\right)-P\left(\hat{p}_{bc}=\frac{j}{L_n}\right)\right|\leq\frac{L_n(1-2p_e)^{b-a}}{1-(1-2p_e)^2}$$

For four sequences S_a, S_b, S_c, and S_d with $a < b < c < d$,

$$(iii) \max_{0\le i\le L_n} \sum_{j=0}^{L_n} \left[P\left(\hat{p}_{bd} = \frac{j}{L_n} \middle| \hat{p}_{ac} = \frac{i}{L_n}\right) - P\left(\hat{p}_{bd} = \frac{j}{L_n}\right)\right] \le \frac{L_n(1-2p_e)^{d-c+b-a}}{1-(1-2p_e)^2}$$

$$(iv) \max_{0\le i\le L_n} \sum_{j=0}^{L_n} \left[P\left(\hat{p}_{bc} = \frac{j}{L_n} \middle| \hat{p}_{ad} = \frac{i}{L_n}\right) - P\left(\hat{p}_{bc} = \frac{j}{L_n}\right)\right] \le \frac{L_n(1-2p_e)^{d-c+b-a}}{1-(1-2p_e)^2}$$

Sketch of Proof of Lemma 11 For a single position r, let $Y_{ij,r} = I_{\{S_{i,r}\neq S_{j,r}\}}$ for all sequences S_i and S_j It follows that, for $i \le j \le k$, $Y_{ij,r}$, $Y_{ik,r}$, and $Y_{jk,r}$ are dependent Bernoulli random variables with respective success probabilities p_{ij}, p_{ik}, and p_{jk}. By Properties 1 and 2 of the total variation distance for functions of independent and identically distributed random variables, if $Y = \sum_{i=1}^n Y_i$ and $X = \sum_{j=1}^n X_j$ for $(X_1, \ldots, X_n) \sim$ Bernoulli(p) and $(Y_1, \ldots, Y_n) \sim$ Bernoulli(q), then

$$\begin{aligned} &\max_{i\in\{0,\ldots,n\}} \sum_{j=0}^{n} \left| P\left(\hat{p} = \frac{j}{n} \middle| \hat{q} = \frac{i}{n}\right) - P\left(\hat{p} = \frac{j}{n}\right)\right| \\ \le(n) &\max_{i\in\{0,1\}} \sum_{j=0}^{1} \left| P\left(X_1 = j \,|Y_1 = i\right) - P\left(X_1 = j\right)\right|. \end{aligned} \tag{A.35}$$

It therefore suffices to establish bounds for a single position r in each case. We provide the details of the calculation for Inequality(i) to illustrate the approach. In this case, we wish to bound

$$\max_i \sum_j \left| P(Y_{ac,r} = j \,|Y_{ab,r} = i\,) - P(Y_{ac,r} = j)\right|$$

for $i, j \in \{0, 1\}$. We first note that

$$P(Y_{ac,r} = 1|Y_{ab,r} = 0) = \frac{(1-p_{a,b})p_{b,c}}{(1-p_{a,b})} = p_{b,c} = \frac{p_{a,b}p_{b,c}}{p_{a,b}} = P(Y_{ac,r} = 0|Y_{ab,r} = 1)$$

$$\text{and } P(Y_{ac,r}=0|Y_{ab,r}=0)=1-p_{b,c}=P(Y_{ac,r}=1|Y_{ab,r}=1).$$

With these calculations,

$$\begin{aligned}&\sum_{j=0}^{1}\left|P(Y_{ac,r}=j\,|Y_{ab,r}=0\,)-P(Y_{ac,r}=j)\right|\\&=|(1-p_{b,c})-(1-p_{a,c})|+|p_{b,c}-p_{a,c}|=2(p_{a,c}-p_{b,c})\end{aligned}\tag{A.36}$$

and

$$\begin{aligned}&\sum_{j=0}^{1}\left|P(Y_{ac,r}=j\,|Y_{ab,r}=1\,)-P(Y_{ac,r}=j)\right|\\&=|p_{b,c}-(1-p_{a,c})|+|(1-p_{b,c}-p_{a,c}|=2(1-p_{a,c}-p_{b,c}).\end{aligned}\tag{A.37}$$

Since $p_{a,c}<\frac{1}{2}$, we have

$$2(p_{a,c}-p_{b,c})\leq 2(1-p_{a,c}-p_{b,c})=(1-2p_e)^{c-b}\left(1+(1-2p_e)^{b-a}\right)\leq 2(1-2p_e)^{c-b}.\tag{A.38}$$

□

Lemma 12

$$\text{(i)}\sum_{k=1}^{g_n+1}\sum_{l=1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right)<\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\frac{g_n^2}{4}+\frac{L_n g_n(1-2p_e)^{\frac{g_n}{2}}}{p_e\left(1-(1-2p_e)^2\right)}\right]$$

$$\text{(ii)}\sum_{k=g_n+2}^{n}\sum_{l=1}^{g_n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right)\leq\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\frac{g_n^2}{4}+\frac{L_n g_n(1-2p_e)^{\frac{g_n}{2}}}{p_e\left(1-(1-2p_e)^2\right)}\right]$$

Proof of Lemma 12 When $1\leq l<k\leq g_n+1$, the distance estimates $\hat{d}_{1,k}$ and $\hat{d}_{l,g_n+1}$ will be highly correlated when l is small (close to 1) and k is close to g_n+1. To account for these terms, we first derive a conservative bound for those terms for which $l<\frac{g_n}{c}$ and $k>\frac{(c-1)g_n}{c}$ for some positive constant

$c < g_n$:

$$\begin{aligned}
&\sum_{k=\frac{(c-1)g_n}{c}+1}^{g_n+1} \sum_{l=1}^{\frac{g_n}{c}-1} \operatorname{Cov}\left(\hat{d}_{1,k}, \hat{d}_{l,g_n+1}\right) \\
&< \left(\frac{g_n}{c}+1\right)\left(\frac{g_n}{c}-1\right)\sqrt{\operatorname{Var}\left(\hat{d}_{1,k}\right)}\sqrt{\operatorname{Var}\left(\hat{d}_{l,g+n+1}\right)} \\
&< \left(\frac{g_n}{c}\right)^2 \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2
\end{aligned} \tag{A.39}$$

where the final inequality stems from the fact that the maximum value for $\hat{d}_{1,k} = d^\star = \frac{1}{2}\ln\left(\frac{L_n}{2}\right)$. For the remaining terms, we have

$$\begin{aligned}
&\sum_{k=1}^{\frac{(c-1)g_n}{c}} \sum_{l=1}^{k-1} \operatorname{Cov}\left(\hat{d}_{1,k}, \hat{d}_{l,g_n+1}\right) + \sum_{k=\frac{(c-1)g_n}{c}+1}^{g_n+1} \sum_{l=\frac{g_n}{c}}^{k-1} \operatorname{Cov}\left(\hat{d}_{1,k}, \hat{d}_{l,g_n+1}\right) \\
&< \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2 \left(\frac{L_n}{1-(1-2p_e)^2}\right) \times \\
&\left[\sum_{k=1}^{\frac{(c-1)g_n}{c}} \sum_{l=1}^{k-1} (1-2p_e)^{g_n-k+l} + \sum_{k=\frac{(c-1)g_n}{c}+1}^{g_n+1} \sum_{l=\frac{g_n}{c}}^{k-1} (1-2p_e)^{g_n-k+l}\right]
\end{aligned} \tag{A.40}$$

by Lemmas 9 and 11. To simplify the preceding inequalities, we bound the sums over l so that the remaining terms may be written as geometric series in k. It follows that

$$\begin{aligned}
\sum_{k=1}^{\frac{(c-1)g_n}{c}} \sum_{l=1}^{k-1} (1-2p_e)^{g_n-k+l} &< \frac{(c-1)g_n}{c} \sum_{k=1}^{\frac{(c-1)g_n}{c}} (1-2p_e)^{g_n-k+1} \\
&= \frac{(c-1)g_n}{c}(1-2p_e)^{\frac{g_n}{c}+1} \sum_{k=0}^{\frac{(c-1)g_n}{c}-1} (1-2p_e)^k \\
&< \frac{(c-1)g_n}{2cp_e}(1-2p_e)^{\frac{g_n}{c}}
\end{aligned} \tag{A.41}$$

and, by the same approach,

$$\sum_{k=\frac{(c-1)g_n}{c}+1}^{g_n+1}\sum_{l=\frac{g_n}{c}}^{k-1}(1-2p_e)^{g_n-k+l} < \frac{(c-1)g_n}{2cp_e}(1-2p_e)^{\frac{g_n}{c}}. \tag{A.42}$$

Combining Inequalities (A.39), (A.41), and (A.42), we have

$$\sum_{k=1}^{g_n+1}\sum_{l=1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right) < \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\left(\frac{g_n}{c}\right)^2+\frac{L_n g_n(1-2p_e)^{\frac{g_n}{c}}}{p_e(1-(1-2p_e)^2)}\right] \tag{A.43}$$

For the second inequality, we consider terms for which $1 \le l < g_n + 1 < k$. We follow the identical approach, first deriving a conservative bound for those terms for which $l \le \frac{g_n}{c}$ and $k \ge \frac{(c-1)g_n}{c}$ and then simplifying the remaining terms to derive the inequality

$$\begin{aligned}
&\sum_{k=g_n+2}^{n}\sum_{l=1}^{g_n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right)\\
&< \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\left(\frac{g_n}{c}\right)^2+\left(\frac{(c-1)g_n}{c}+g_n\right)\frac{L_n(1-2p_e)^{\frac{g_n}{c}}}{2p_e\left(1-(1-2p_e)^2\right)}\right]\\
&< \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\left(\frac{g_n}{c}\right)^2+\frac{L_n g_n(1-2p_e)^{\frac{g_n}{c}}}{p_e\left(1-(1-2p_e)^2\right)}\right].
\end{aligned} \tag{A.44}$$

Since Inequalities A.43 and A.44 hold for any integer $c \in [2, g_n - 1]$, we choose $c = 2$ for convenience to derive the final results. □

Lemma 13

$$\sum_{k=g_n+2}^{n}\sum_{l=g_n+1}^{n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right) \le \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\frac{2L_n(n-g_n)^2(1-2p_e)^{g_n}}{1-(1-2p_e)^2}$$

Proof of Lemma 13 We first consider the terms in the summation for which $1 < g_n + 1 < k \le l$. In this region, the distances d_{1k} and $d_{g_n+1,l}$ partially

overlap, and some correlation between the distance estimates is expected. Applications of Lemmas 9 and 11 give the inequality

$$\begin{aligned}\sum_{k=g_n+2}^{n}\sum_{l=k}^{n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right) \leq & \sum_{k=g_n+2}^{n}\sum_{l=k}^{n}\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\frac{L_n(1-2p_e)^{l-k+g_n}}{1-(1-2p_e)^2}\\ \leq & (n-g_n)^2\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\frac{L_n(1-2p_e)^{g_n}}{1-(1-2p_e)^2}.\end{aligned} \tag{A.45}$$

We now consider the terms with $1 < g_n + 1 \leq l < k$. In this region, the correlations between estimates $\hat{d}_{1,k}$ and $\hat{d}_{g_n+1,l}$ are weakened by the large distance between sequences S_1 and S_{g_n}. Again applying Lemmas 9 and 11 and bounding as above, we derive the inequality

$$\sum_{k=g_n+2}^{n}\sum_{l=g_n+1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right) < (n-g_n)^2\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\frac{L_n(1-2p_e)^{g_n}}{1-(1-2p_e)^2}. \tag{A.46}$$

□

Proof of Theorem 3 From Lemma 9,

$$\begin{aligned}&\sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{Cov}\left(\hat{d}_{1,k}-\hat{d}_{g_n,k},\hat{d}_{2l}-\hat{d}_{g_n+1,l}\right)\\ \geq &-\sum_{k=1}^{n}\sum_{l=1}^{n}\left[\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right)+\mathrm{Cov}\left(\hat{d}_{g_n,k},\hat{d}_{2l}\right)\right].\end{aligned} \tag{A.47}$$

The proof then consists of deriving the following bounds:

$$\sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n+1,l}\right) \leq \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\frac{g_n^2}{2}+o\left(n^{-1}\right)\right] \tag{A.48}$$

$$\text{and } \sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{Cov}\left(\hat{d}_{g_n,k},\hat{d}_{2l}\right) \leq \left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\frac{g_n^2}{2}+o\left(n^{-1}\right)\right]. \tag{A.49}$$

To establish the first inequality, we divide the sum into sections and bound each case. Because the derivation of Inequality (A.49) is nearly identical to that of Inequality (A.48), the result is stated without proof.

In the first region, we consider terms with either $1 \leq k \leq g_n + 1 \leq l$ or $1 \leq k \leq l \leq g_n + 1$. The distance estimates $\hat{d}_{1,k}$ and $\hat{d}_{g_n+1,l}$ are independent in either case, and thus

$$\sum_{k=1}^{g_n+1} \sum_{l=k}^{n} \mathrm{Cov}\left(\hat{d}_{1,k}, \hat{d}_{g_n+1,l}\right) = 0 \tag{A.50}$$

We bound the remaining regions in Lemmas 12 and 13. These results, in combination, cover all of the n^2 terms in the sum $\sum_{k=1}^{n} \sum_{l=1}^{n} \mathrm{Cov}\left(\hat{d}_{1,k}, \hat{d}_{g_n+1,l}\right)$, and so we derive the overall bound:

$$\begin{aligned}
&\sum_{k=1}^{n} \sum_{l=1}^{n} \mathrm{Cov}\left(\hat{d}_{1,k}, \hat{d}_{g_n+1,l}\right) \\
&< \left(\frac{1}{2} \ln\left(\frac{L_n}{2}\right)\right)^2 \left[\frac{g_n^2}{2} + \frac{2L_n g_n (1-2p_e)^{\frac{g_n}{2}}}{p_e\left(1-(1-2p_e)^2\right)} + \frac{2L_n (n-g_n)^2 (1-2p_e)^{g_n}}{1-(1-2p_e)^2}\right] \\
&= \left(\frac{1}{2} \ln\left(\frac{L_n}{2}\right)\right)^2 \left[\frac{g_n^2}{2} + o\left(n^{-1}\right)\right]
\end{aligned}$$

for $g_n = n^\gamma$ with $\gamma > 0$. $\square$

Sketch of Proof of Corollary 4 We summarize the result for the first inequality. By Lemma 9,

$$\sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \mathrm{Cov}\left(\Delta_k, \Delta_l\right) \geq -\sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \mathrm{Cov}\left[\left(\hat{d}_{1,k}, \hat{d}_{g_n,l}\right) + \mathrm{Cov}\left(\hat{d}_{g_n,k}, \hat{d}_{1,l}\right)\right] \tag{A.51}$$

For this first term,

$$\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)=\sum_{k=g_n+1}^{n-1}\sum_{l=k+1}^{n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right) \tag{A.52}$$

by the independence of $\hat{d}_{1,k}$ and $\hat{d}_{g_n,l}$ for $k < g_n < l$ and $k < l < g_n$. And following the proof of Lemma 13, we find that

$$\sum_{k=g_n+1}^{n-1}\sum_{l=k+1}^{n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)\leq(n-g_n)^2\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\frac{L_n(1-2p_e)^{g_n}}{1-(1-2p_e)^2}. \tag{A.53}$$

For the second term, we note that

$$\begin{aligned}&\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\mathrm{Cov}\left(\hat{d}_{g_n,k},\hat{d}_{1,l}\right)=\sum_{k=2}^{n}\sum_{l=1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)\\&=\sum_{k=2}^{g_n}\sum_{l=1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)+\sum_{k=g_n+1}^{n}\sum_{l=1}^{g_n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)+\sum_{k=g_n+1}^{n}\sum_{l=g_n+1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right).\end{aligned} \tag{A.54}$$

For the first two summations in Equation (A.54), we follow the proof of Lemma 12 to establish that

$$\begin{aligned}&\sum_{k=2}^{g_n}\sum_{l=1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)+\sum_{k=g_n+1}^{n}\sum_{l=1}^{g_n}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)\\&\leq\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\frac{g_n^2}{2}+\frac{2L_ng_n(1-2p_e)^{\frac{g_n}{2}}}{p_e\left(1-(1-2p_e)^2\right)}\right],\end{aligned} \tag{A.55}$$

and for the third summation, we follow the proof of Lemma 13 to obtain

$$\sum_{k=g_n+1}^{n}\sum_{l=g_n+1}^{k-1}\mathrm{Cov}\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)\leq(n-g_n)^2\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\frac{L_n(1-2p_e)^{g_n}}{1-(1-2p_e)^2}. \tag{A.56}$$

Combining these results, we have

$$\begin{aligned}
&\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\mathrm{Cov}\left[\left(\hat{d}_{1,k},\hat{d}_{g_n,l}\right)+\mathrm{Cov}\left(\hat{d}_{g_n,k},\hat{d}_{1,l}\right)\right] \\
&<\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\frac{g_n^2}{2}+\frac{2L_ng_n(1-2p_e)^{\frac{g_n}{2}}}{p_e\left(1-(1-2p_e)^2\right)}+\frac{2L_n(n-g_n)^2(1-2p_e)^{g_n}}{1-(1-2p_e)^2}\right] \\
&=\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[\frac{g_n^2}{2}+O\left(n^{-1}\right)\right] \qquad \text{(A.57)}
\end{aligned}$$

□

Lemmas 14 and 15 are employed in the Proof of Theorem 5.

Lemma 14

$$\mathrm{Var}\left(D_n\right)=\left(\frac{1}{n-2}\right)^2\mathrm{Var}\left(\left(\hat{d}_{1.}-\hat{d}_{g_n.}\right)+\left(\hat{d}_{2.}-\hat{d}_{g_n+1.}\right)\right)+O\left(n^{\frac{-(s-1)}{2}}\ln(n)\right).$$

Proof of Lemma 14 Beginning with the variance of $\left(\hat{d}_{g_n,g_n+1}-\hat{d}_{12}\right)$, we see that, because the random variables $\hat{d}_{g_n,g_n+1}$ and d_{12} are independent and identically distributed,

$$\mathrm{Var}\left(\hat{d}_{g_n,g_n+1}-\hat{d}_{12}\right)=\mathrm{Var}\left(\hat{d}_{g_n,g_n+1}\right)+\mathrm{Var}\left(\hat{d}_{12}\right)=2\mathrm{Var}\left(\hat{d}_{12}\right).$$

Define X_{12} to be a random variable which counts the number of observed differences between sequences 1 and 2. Then X_{12} follows a Binomial distribution with size L_n and probability p_{12}, and, for large values of L_n, the proportion of differences $\hat{p}_{12}$ will be approximately normally distributed with mean p_{12} and variance $\frac{p_{12}(1-p_{12})}{L_n}$. Because $p_{12}=p_e$ and $L_n=n^s$, it is clear that $\mathrm{Var}\left(\hat{p}_{12}\right)=\frac{1-(1-2p_e)^2}{4n^s}\to 0$ as $n\to\infty$ for any constant value p_e. Since $\hat{d}_{12}=-\frac{1}{2}\ln\left(1-2\hat{p}_{12}\right)$, by straightforward Taylor series calculations in the

neighborhood of p_{12} we have $\mathrm{Var}\left(\hat{d}_{12}\right) \sim \frac{1-(1-2p_e)^2}{4n^s(1-2p_e)^2} \to 0$, and it follows that $\mathrm{Var}\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}\right) = O\left(n^{-s}\right)$.

For the covariance terms, we have

$$\left|\mathrm{Cov}\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}, \hat{d}_{1.} - \hat{d}_{g_n.}\right)\right| \leq \sqrt{\mathrm{Var}\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}\right)}\sqrt{\mathrm{Var}\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right)} \tag{A.58}$$

and

$$\left|\mathrm{Cov}\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}, \hat{d}_{2.} - \hat{d}_{g_n+1.}\right)\right| \leq \sqrt{\mathrm{Var}\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}\right)}\sqrt{\mathrm{Var}\left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)}. \tag{A.59}$$

Because $\mathrm{Var}\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}\right) = O(n^{-s})$, we need only show that the variance of $\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right)$ and $\left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)$ is not growing at a rate greater than or equal n^s to establish the overall convergence of the covariance terms. To bound $\mathrm{Var}\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right)$, we see that

$$\begin{aligned}\mathrm{Var}\left(\sum_{k=1}^{n}\Delta_k\right) &= \sum_{k=1}^{n}\mathrm{Var}\left(\hat{d}_{1,k}\right) + \sum_{k=1}^{n}\mathrm{Var}\left(\hat{d}_{g_n,k}\right) - 2\sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{Cov}\left(\hat{d}_{1,k}, \hat{d}_{g_n,l}\right) \\ &\leq \sum_{k=1}^{n}\mathrm{Var}\left(\hat{d}_{1,k}\right) + \sum_{k=1}^{n}\mathrm{Var}\left(\hat{d}_{g_n,k}\right)\end{aligned}$$

by the positivity of the covariance of all pairs of distances as established in Lemma 9. Bounding each variance term by the largest possible value $\left(d^\star\right)^2$, we have

$$\mathrm{Var}\left(\sum_{k=1}^{n}\Delta_k\right) \leq 2n\left(d^\star\right)^2 = \frac{n}{2}\left(\ln\left(\frac{L_n}{2}\right)\right)^2 < \frac{s^2 n}{2}\left(\ln(n)\right)^2 = O(n(\ln(n))^2),$$

and it follows that

$$\left|\mathrm{Cov}\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}, \hat{d}_{1.} - \hat{d}_{g_n.}\right)\right| = \left|\sqrt{O\left(n^{-s}\right) \times O(n(\ln(n))^2)}\right| = O\left(n^{\frac{-(s-1)}{2}} \ln(n)\right).$$

□

Lemma 15

$$\mathrm{Var}\left(\sum_{k=1}^{n} \Delta_k\right) + \mathrm{Var}\left(\sum_{k=1}^{n} \Delta'_k\right) > 2\left(\ln(n)\right)^2 \left[(n - (g_n + b_n))c_{\beta,s,n} - \frac{g_n^2 s^2}{4} - o\left(n^{-1}\right)\right].$$

Proof of Lemma 15 Expanding the first term in the expression, we have

$$\mathrm{Var}\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right) = \sum_{k=1}^{n} \mathrm{Var}\left(\Delta_k\right) + 2\sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \mathrm{Cov}\left(\Delta_k, \Delta_l\right) \tag{A.60}$$

Beginning with the variance terms, we consider only those values of k for which $k \geq g_n + b_n$, where $b_n = n^\beta$ for some $\beta \in \left(0, \frac{1}{2}\right)$. In this region, both the distance estimates $\hat{d}_{1,k}$ and $\hat{d}_{g_n,k}$ are likely to be "corrected" to the value $d^\star$ with nearly equal probability, since, for sequences of polynomial length, the true mutation probabilities $\hat{p}_{1,k}$ and $\hat{p}_{g_n,k}$ will both be greater than $p^\star = \left(\frac{1}{2} - \frac{1}{L_n}\right)$. And, because g_n is large, the random variables $\hat{p}_{1,k}$ and $\hat{p}_{g_n,k}$ are nearly independent. Thus, for any pair of distance estimates $\left\{\hat{d}_{1,k}, \hat{d}_{g_n,k}\right\}$, the probability that either estimate is corrected to the value $d^\star$ will approach $\frac{1}{2}$ for large values of k. By Theorem 2, this behavior results in the inequality

$$\mathrm{Var}(\hat{d}_{1,k} - \hat{d}_{g_n,k}) \geq \left(\frac{1}{4} - \frac{\delta_{g_n k}}{2} - o\left(n^{-1}\right)\right)\left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n}\right)^2 (\ln(n))^2,$$

where $\delta_{g_n k} = P\left(\hat{p}_{g_n,k} < \frac{1}{2}\right) - \frac{1}{2} \to 0$ as $k \to \infty$. Applying this inequality to the terms for which $k > g_n + b_n$, we have

$$\begin{aligned}
&\sum_{k=g_n+b_n}^{n} \mathrm{Var}\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) \\
>&(n - (g_n + b_n)) \left(\frac{1}{4} - \frac{\delta_{g_n,g_n+b_n}}{2} - o\left(n^{-1}\right)\right) \left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n}\right)^2 (\ln(n))^2 \\
=&(n - (g_n + b_n)) c_{\beta,s,n} \left(\ln(n)\right)^2 \qquad \text{(A.61)}
\end{aligned}$$

where $c_{\beta,s,n} \to \left(\frac{1}{4}\right)\left(\frac{s}{4} - \frac{1}{2}\right)^2$ as $n \to \infty$. And for the covariance terms in Equation(A.60),

$$\sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \mathrm{Cov}\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}, \hat{d}_{1l} - \hat{d}_{g_n,l}\right) > -\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2 \left[\frac{g_n^2}{2} + o\left(n^{-1}\right)\right]$$

by Corollary 4. For $\mathrm{Var}\left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)$, we note that, for all $k > g_n + 1$, $\mathrm{Var}\left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right) = \mathrm{Var}\left(\hat{d}_{1,k-1} - \hat{d}_{g_n,k-1}\right)$. We therefore bound the sum as for the previous term and again apply Corollary 4, deriving the identical bound

$$\mathrm{Var}\left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right) > \left(\ln(n)\right)^2 \left[(n - (g_n + b_n)) c_{\beta,s,n} - \frac{g_n^2 s^2}{4} - o\left(n^{-1}\right)\right]. \qquad \text{(A.62)}$$

□

Proof of Theorem 5 By definition,

$$
\begin{aligned}
&\operatorname{Var}\left(\hat{D}_{g_n g_n+1}-\hat{D}_{12}\right)\\
=&\operatorname{Var}\left(\left(\hat{d}_{g_n,g_n+1}-\hat{d}_{12}\right)+\frac{1}{n-2}\left(\hat{d}_{1.}-\hat{d}_{g_n.}\right)+\frac{1}{n-2}\left(\hat{d}_{2.}-\hat{d}_{g_n+1.}\right)\right)\\
=&\operatorname{Var}\left(\hat{d}_{g_n,g_n+1}-\hat{d}_{12}\right)+\left(\frac{1}{n-2}\right)^2\left[\operatorname{Var}\left(\hat{d}_{1.}-\hat{d}_{g_n.}\right)+\operatorname{Var}\left(\hat{d}_{2.}-\hat{d}_{g_n+1.}\right)\right]\\
&+\left(\frac{2}{n-2}\right)\left[\operatorname{Cov}\left(\hat{d}_{g_n,g_n+1}-\hat{d}_{12},\hat{d}_{1.}-\hat{d}_{g_n.}\right)+\operatorname{Cov}\left(\hat{d}_{g_n,g_n+1}-\hat{d}_{12},\hat{d}_{2.}-\hat{d}_{g_n+1.}\right)\right]\\
&+2\left(\frac{1}{n-2}\right)^2\operatorname{Cov}\left(\hat{d}_{1.}-\hat{d}_{g_n.},\hat{d}_{2.}-\hat{d}_{g_n+1.}\right).
\end{aligned}
$$

To analyze this expression, we first note that, by Lemma 14, the terms involving $\left(\hat{d}_{g_n,g_n+1}-\hat{d}_{12}\right)$ are negligible for sequences of polynomial length. Focusing on the remaining terms, we employ the lower bounds for $\operatorname{Var}\left(\hat{d}_{1.}-\hat{d}_{g_n.}\right)$ and $\operatorname{Var}\left(\hat{d}_{2.}-\hat{d}_{g_n+1.}\right)$ obtained in Lemma 15. And by Theorem 3,

$$
\operatorname{Cov}\left(\hat{d}_{1.}-\hat{d}_{g_n.},\hat{d}_{2.}-\hat{d}_{g_n+1.}\right)\geq-\left(\frac{1}{2}\ln\left(\frac{L_n}{2}\right)\right)^2\left[g_n^2+o\left(n^{-1}\right)\right].
$$

We combine this inequality with the results from Lemmas 14 and 15 to establish the overall variance bound.

□

A.1 Total Variation Distance Results

The following properties are easily verified, and are therefore stated without proof.

Property 1 For two sequences of independent and identically distributed random variables $Y_1, Y_2, \ldots, Y_k$ and $Z_1, Z_2, \ldots, Z_k$,

$$\mathrm{TV}\left((Y_1, Y_2, \ldots, Y_k), (Z_1, Z_2, \ldots, Z_k)\right) \leq (k)\mathrm{TV}\left(Y_1, Z_1\right).$$

Property 2 For any two random vectors $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)$ and any function f,

$$\mathrm{TV}\left(f(\mathbf{Y}), f(\mathbf{Z})\right) \leq \mathrm{TV}(\mathbf{Y}, \mathbf{Z}).$$

Acknowledgements

We thank the reviewers for their helpful comments and suggestions.

References

[1] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular Biology and Evolution 4 (1987) 406–425.

[2] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis, Cambridge University Press, 1998.

[3] K. Atteson, The perfomance of neighbor-joining methods of phylogenetic reconstruction, Algorithmica 25 (1999) 251–278.

[4] D. Huson, S. Nettles, T. Warnow, Disk-covering, a fast-converging

method for phylogenetic tree reconstruction, Journal of Computational Biology 6 (1999) 369–386.

[5] P. Erdős, M. Steel, L. Székély, T. Warnow, A few logs suffice to build almost all tress – I, Random Structures and Algorithms 14 (1997) 153–184.

[6] D. Huson, S. Nettles, L. Parida, T. Warnow, S. Yooseph, A divide-and-conquer approach to tree reconstruction, in: Workshop on Algorithms and Experiments, (ALEX98), Trento, Italy, 1998.

[7] T. Warnow, B. Moret, K. St. John, Absolute convergence: True trees from short sequences, in: IEEE Symposium on Discrete Algorithms (SODA 01), 2001, pp. 186–195.

[8] L. Nakhleh, U. Roshan, K. St. John, J. Sun, T. Warnow, Designing fast converging phylogenetic methods, in: ISMB (Supplement of Bioinformatics), 2001, pp. 190–198.

[9] L. Nakhleh, B. Moret, U. Roshan, K. St. John, J. Sun, T. Warnow, The accuracy of fast phylogenetic methods for large datasets, in: Proc. 7th Pacific Symp. on Biocomputing (PSB 2002), 2002, pp. 211–222.

[10] K. St. John, T. Warnow, B. Moret, L. Vawter, Performance study of phylogenetic methods: (unweighted) quartet methods and N-J, in: IEEE Symposium on Discrete Algorithms (SODA 01), 2001, pp. 196–205.

[11] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, Journal of Molecular Biology 16 (1980) 111–120.

[12] M. Hasegawa, H. Kishino, T. Yano, Dating the human-ape splitting by a molecular clock of mitochondrial DNA, Journal of Molecular Biology 22 (1985) 160–174.

[13] M. Nei, T. Gojobori, Simple methods for estimating the numbers of syn-

onymous and nonsynonymous nucleotide substitutions, Molecular Biology and Evolution 3 (5) (1986) 418–26.

[14] W. Li, C. Wu, C. Luo, A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes, Molecular Biology and Evolution 2 (2) (1985) 150–74.

[15] D. Hillis, J. Bull, M. White, R. Badgett, I. Molineux, Experimental phylogenetics: Generation of a known phylogeny, Science 255 (1992) 589–592.

[16] T. Buckley, Model misspecification and probabilistic tests of topology: Evidence from empirical data sets, Systematic Biology 51 (3) (2002) 509–523.

[17] G. Sanson, S. Kawashita, A. Brunstein, M. Briones, Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions, Molecular Biology and Evolution 19 (2) (2002) 170–178.

[18] D. Posada, T. Buckley, Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests, Systematic Biology 53 (5) (2004) 793–808.

[19] T. H. Jukes, C. Cantor, Mammalian Protein Metabolism, Academic Press, 1969, pp. 21–132.

[20] D. Huson, K. Smith, T. Warnow, Estimating large distances in phylogenetic reconstruction, in: Lecture Notes in Computer Science (Algorithm Engineering) 1668, Vol. 1668, London, UK, 1999, pp. 271–285, proceedings from the 3rd International Workshop, WAE'99.

[21] D. L. Swofford, G. Olse, P. Waddell, D. Hillis, Phylogenetic inference, in: D. Hillis, C. Moritz, B. Mable (Eds.), Molecular Systematics, 2nd Edition,

Sinauer Associates, Inc., 1996, Ch. 11, pp. 407–514.

[22] J. Cavender, Taxonomy with confidence, Mathematical Biosciences 40 (1978) 271–280.

[23] J. S. Farris, A probability model for inferring evolutionary trees, Systematic Zoology 22 (1973) 250–256.

[24] J. Huelsenbeck, Performance of phylogenetic methods in simulation, Systematic Biology 44 (1) (1995) 17–48.

[25] K. Takahashi, M. Nei, Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used, Molecular Biology and Evolution 17 (2000) 1251–1258.

[26] M. S. Rosenberg, S. Kumar, Phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well, Molecular Biology and Evolution 18 (2001) 1823–1827.

[27] K. Tamura, M. Nei, S. Kumar, Prospects for inferring very large phylogenies by using the neighbor-joining method, Proc Natl Acad Sci USA 101 (30) (2004) 11030–5.

[28] E. Lehmann, Some concepts of dependence, Annals of Mathematical Statistics 37 (1966) 1137–1153.

Figure Captions

Figure 1: (a) A general n-taxa caterpillar tree. There are $n-2$ internal nodes, each represented by a "dot" in the figure, and all edges have equal length d_e. (b) The "legless" caterpillar tree considered in the analysis. The taxa are sequentially connected by edges of equal length d_e.

Figure 2: Reconstructing a 4-leaf caterpillar. Two of the four correct paths, shown on the left and right, begin by joining leaves 1 and 2. At each step, leaves and internal nodes included in the N-J distance calculations are enclosed in boxes.

Figures

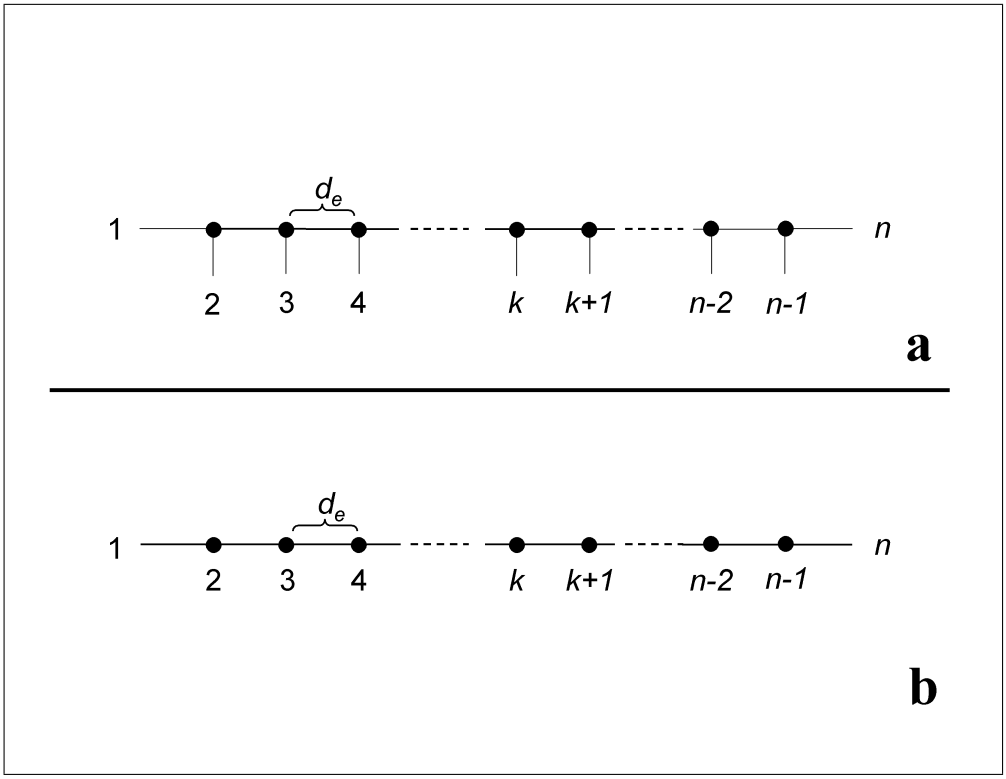

Fig. 1.

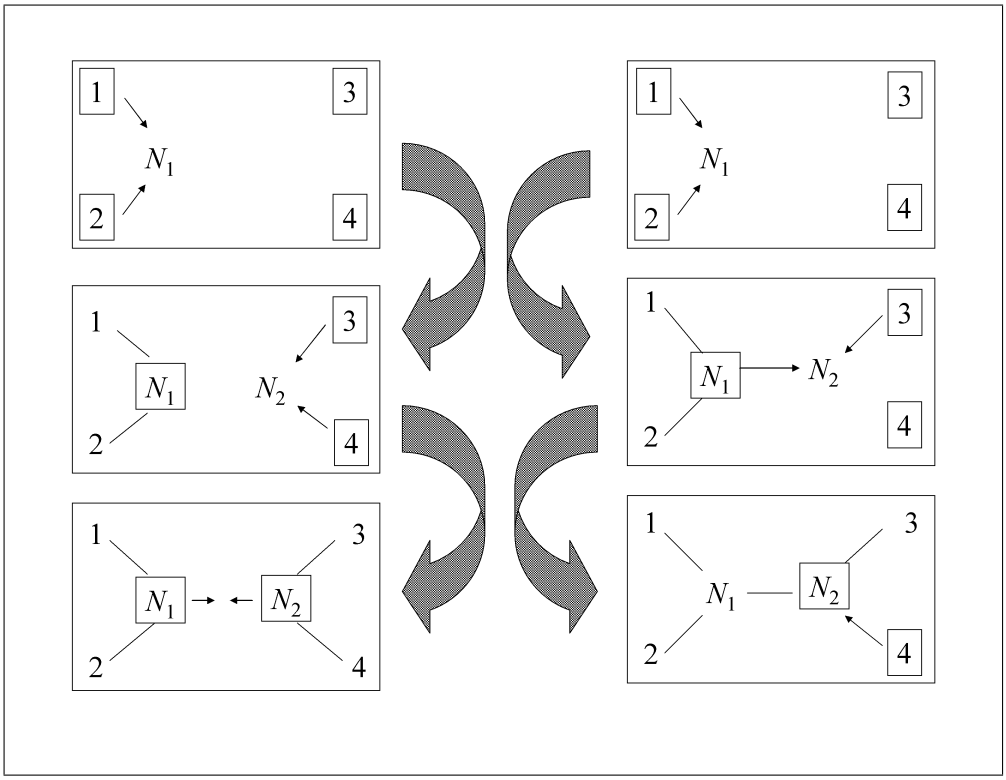

Fig. 2.