Molecular Phylogenetics and Evolution 93 (2015) 289–295

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Twisted trees and inconsistency of tree estimation when gaps are treated as missing data – The impact of model mis-specification in distance corrections ☆

Emily Jane McTavish [a,b,*], Mike Steel [c], Mark T. Holder [a,b]

[a] Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[b] Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA
[c] Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 28 April 2015
Revised 9 July 2015
Accepted 21 July 2015
Available online 6 August 2015

Keywords:
Phylogenetics
Distance methods
Invariant sites
Insertion
Deletion
Gaps as missing data

ABSTRACT

Statistically consistent estimation of phylogenetic trees or gene trees is possible if pairwise sequence dissimilarities can be converted to a set of distances that are proportional to the true evolutionary distances. Susko et al. (2004) reported some strikingly broad results about the forms of inconsistency in tree estimation that can arise if corrected distances are not proportional to the true distances. They showed that if the corrected distance is a concave function of the true distance, then inconsistency due to long branch attraction will occur. If these functions are convex, then two "long branch repulsion" trees will be preferred over the true tree – though these two incorrect trees are expected to be tied as the preferred true. Here we extend their results, and demonstrate the existence of a tree shape (which we refer to as a "twisted Farris-zone" tree) for which a single incorrect tree topology will be guaranteed to be preferred if the corrected distance function is convex. We also report that the standard practice of treating gaps in sequence alignments as missing data is sufficient to produce non-linear corrected distance functions if the substitution process is not independent of the insertion/deletion process. Taken together, these results imply inconsistent tree inference under mild conditions. For example, if some positions in a sequence are constrained to be free of substitutions and insertion/deletion events while the remaining sites evolve with independent substitutions and insertion/deletion events, then the distances obtained by treating gaps as missing data can support an incorrect tree topology even given an unlimited amount of data.

1. Introduction

Distance-based methods are fast and statistically consistent estimators of tree topology if the input distances converge (with increasing data) to values that are proportional to the evolutionary distance between tips. An evolutionary distance is the number of substitution events that have occurred along the path separating two tips. Typically a distance correction procedure is applied to the observed sequence differences to obtain a more accurate estimate of the evolutionary distance between pairs of sequences. However, in many cases it is not possible to correctly account for the evolutionary processes which generated the data. In other words, it is not always possible to accurately estimate the evolutionary distance for pairwise measurements of dissimilarity.

In a pioneering paper, Susko et al. (2004) showed how model misspecification can lead to transformed evolutionary distances that are non-linear with respect to evolutionary distance (i.e. concave or convex), and for which there are regions of tree space for which neighbor joining will be inconsistent. We extend this result further (Theorem 1 in Appendix A) by showing how virtually all misspecified correction functions lead to (strong) inconsistency (an incorrect tree will be unambiguously favored by neighbor-joining). A main focus of this paper involves a particular study of model misspecification in distance corrections that treats gaps as missing data.

2. Model

For variants of the simplest model of sequence evolution (Jukes and Cantor, 1969), all nucleotides are equally exchangeable and the simple proportion of sites that differ, the *p*-distance, is a sufficient statistic for estimating an evolutionary distance. Under such a

☆ This paper was edited by the Associate Editor Scott Edwards.

* Corresponding author at: Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas, USA.
E-mail address: ejmctavish@gmail.com (E.J. McTavish).

http://dx.doi.org/10.1016/j.ympev.2015.07.027
1055-7903/

model, M_g, the expected p-distance between a pair of taxa is a function of the evolutionary distance (path length in the tree) t between the taxa, that is, we have $\mathbb{E}_g[p] = g(t)$, where the function g is a monotonically (strictly) increasing function of t which is analytic (i.e. has a power series expansion, and so derivatives exist of all orders) and satisfies $g(0) = 0$. For example, for the Jukes–Cantor model we have $g(t) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}t}\right)$. If the distances are corrected under a (possibly different), fully exchangeable model, M_f, then the estimated evolutionary distance $\hat{t}$ is usually computed from the p-distance by using the 'plug-in' formula $\hat{t} = f^{-1}(p)$.

Thus, for any generating model for which p converges in probability towards its expected value $\mathbb{E}_g[p] = g(t)$ (e.g. i.i.d. site substitution models) the estimated evolutionary distance $\hat{t}$ will converge towards $\bar{t} = h(t)$, where $h(t) = f^{-1}(g(t))$. Note here that both p and $\hat{t}$ are random variables, while $\bar{t}$ is simply a function of t. Notice that this 'transformed' evolutionary distance $\bar{t}$ is not exactly the expected value of $\hat{t}$, even when $f = g$ (Tajima, 1993), since the expectation of a non-linear function of random variable is not generally equal to the function evaluated at the expected value of that variable. Nevertheless, for any i.i.d. site substitution model, the difference between $\bar{t}$ and the expected value of $\hat{t}$ decays towards zero as the sequence length grows.

Notice also that when $f = g$ (i.e. the correction model matches the generating model) then $\bar{t} = t$. However, in general, $\bar{t}$ need not be equal to t, except when $t = 0$. For example, if the generating model is the Jukes–Cantor model with some form of among-site rate heterogeneity and the correcting model that does not assume the same form of rate heterogeneity then $\bar{t}$ can depend on t in a quite non-linear way (Soubrier et al., 2012).

In this paper we are interested in determining when the transformed evolutionary distances $\bar{t}$ will favor a different tree to the tree generating the data. In particular, we explore an example of how ignoring the process of insertion and deletion (referred to jointly as indels hereafter) can lead to statistical inconsistency in an otherwise correctly modeled substitution process. Inconsistency occurs in this case even when the alignment of residues is correct.

Susko et al. (2004) studied general properties of $\bar{t}$ as a function of t. If this function is linear (i.e. when the correction model matches the generating model up to a constant factor) then distance-based tree estimation will be statistically consistent. If the function is concave, inference can be inconsistent and positively misleading due to long branch attraction. They also show that if the function is convex, two long branch repulsion trees are expected to be equally preferred over the correct tree. In Appendix A we establish a more general result: outside of the special case where the correcting generating model matches the generating model up to a constant factor, there will always exist tree shapes for which neighbor-joining will estimate a single incorrect tree from $\bar{t}$. The tree shapes used to demonstrate this result are the familiar Felsenstein-zone tree (Fig. 1A; Felsenstein, 1978) and a tree that we refer to as the "twisted Farris-zone" tree (Fig. 1B). "Farris-zone" tree is used to refer to tree shapes that exhibit long branch repulsion under certain conditions of model violation, and this asymmetrical ("twisted") variant has branch lengths which will result in a single incorrect tree topology being preferred if the corrected distance function is convex.

2.1. The gaps as missing data convention

It is common practice to treat a gap in a sequence as missing data in phylogenetic estimation based on distances, parsimony scores or likelihoods. In the context of a pairwise distance

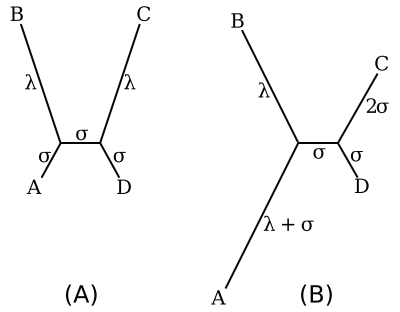

Fig. 1. (A) The Felsenstein-zone tree with branch lengths used in the proof of Lemma 3; (B) The "twisted Farris-zone" tree used in the proof of Lemma 4.

calculation, this treatment means that positions with a gap in either sequence are disregarded because they cannot be counted as either a similarity or a difference. Omitting indels from distance corrections obviously forfeits the opportunity for learning about the evolutionary distance from insertions and deletion events. However, one may hope that treating sites with gaps as missing data would not perturb a distance estimate that relies solely on substitution events. If the substitution and indel processes are completely independent, and have the same stationary nucleotide frequencies, this is the case.

Consider the case of sequences that are generated by: a time-reversible stochastic process of insertions and deletion, and a model of substitutions for which there is a statistically consistent distance correction. If the alignment is known without error, then the only effect of the indel process is to introduce a fraction of sites, z, for which one sequence lacks a residue and the other sequence has a residue. These are the gapped positions in a pairwise alignment. Note that the presence of gap in a column in the alignment is not handled by deleting the column. The gap only affects pairwise comparisons involving a sequence which contains a gap. A full description for z for a full alignment would require some additional notation to indicate which sequences are being compared. Our argument below applies to any pairwise distance, so we simply use $z(t)$ to describe the expected proportion of gapped position in any pairwise distance for sequences separated by path length, t.

The fraction of gapped positions will be a function of the evolutionary distance with: $z(0) = 0$ because at no distance there are no opportunities for indels, and $z(t) < 1$ for all t. The latter property can be seen by treating one of the two sequences as if it were the ancestral sequence. This is permissible because we have assumed that the indel process is time reversible. The probability of a residue surviving from the ancestral sequence to the descendant sequence is described by an exponential function with rate parameter controlled by the rate of deletions. This probability remains > 0 for all values of the evolutionary distance, hence there is a non-zero probability of an ungapped position, and $z(t)$ cannot equal 1.

In a typical consistency proof, we consider sequences of ever increasing length. We note that indel models (e.g. the TKF91 model; Thorne et al. (1991)) imply a equilibrium sequence length. Here we discuss statistical consistency by considering what happens as the number of loci increases without bound, but the equilibrium length of each locus is determined by the parameters of the indel model. Hence the total sequence length approaches infinity, while it is still appropriate to describe the sequence as being generated by the indel process.

For the standard substitution models, we can consistently estimate the distance if the indel process has insertion and deletion rates of 0. In this case there are no gapped columns and $z(t) = 0$. In the more general case, if we only consider site patterns in which no gaps occur, the probability of a site pattern s for branch length t

is $\Pr(s|t) = (1 - z(t))\Pr'(s|t)$ where $\Pr'(s|t)$ is the usual site pattern probability (when we have no missing data caused by gaps), and $(1 - z(t))$ is the probability of not containing a gap. The multiplication of the probability of not containing a gap by the probability of the site pattern given the branch length is valid whenever the substitution and indel process are independent. We can see that calculating the probability of each ungapped site pattern by using the fraction of ungapped sites that display the pattern will result in $\Pr'(s|t)$ because this will constitute dividing the probability of each pattern by $(1 - z(t))$. Thus the spectrum of ungapped pattern frequencies will converge to exactly the same frequencies of the patterns when there are no indel events. If the insertion and deletion process and the substitution process have different equilibrium nucleotide frequencies, or if the probability of a deletion is affected by the nucleotide base at a site, this consistency may not hold.

Thus, if the indel process and substitution process are independent, treating gaps as missing data will not cause statistical inconsistency of distance-based tree inference. Note that this result does not contradict the proof by Warnow (2012) that treating gaps as missing data can lead to inconsistency in maximum likelihood. Her proof focuses on the maximum likelihood criterion and is restricted to the case in which internal branch lengths for the substitution process are equal to 0 (there are no substitution events). Internal branch lengths of 0 lead to inconsistency without the complication of an indel process. Our result applies to cases in which the tree method is capable of consistently estimating the tree if there are no indels.

2.2. Cases in which indel processes and substitution process are not independent

If the occurrence of an indel affects the probability of a substitution, then the previous argument does not hold. In fact, we cannot use the argument above under any violation of the independence assumption. For example, if some subset of sites is constrained by evolution and thus free of both substitutions and indels, then it is possible for the gaps-as-missing-data convention to lead us to the wrong tree. In such cases, the gapped sites are a biased sample with respect to the substitution process. See Roure et al. (2013) for a discussion of other contexts in which non-random patterns of missing data perturb phylogenetic estimation. Specifically, if the distribution of missing sites is not independent of the evolutionary rates at those sites this bias can lead to problems in phylogenetic reconstruction (Grievink et al., 2013; Roure et al., 2013).

2.3. Paired invariants model

Consider the case of sequences being generated under the TKF91 (Thorne et al., 1991) indel model and the Jukes–Cantor (JC) (Jukes and Cantor, 1969) substitution model, but with invariant sites. In particular, let the "paired invariant sites" model refer to the case in which some fraction of sites are free from both indels and substitutions and the other parts of the sequence are described by the TKF91 and JC models. In terms of the formalism of the TKF91 model, this would require that each invariant site which is followed by a region that is free to vary is considered to have a new "immortal link" to the right of the invariant site. We consider the case in which alignment is known without error.

Let p_{inv} denote the expected proportion sites in a sequence which are invariant with respect to indels and substitutions. In the TKF model single nucleotide insertions and deletions can occur at any site in the alignment (Thorne et al., 1991). Under the TKF model, at equilibrium the expected rate of insertions per locus is equal to the expected rate of deletions per locus. The TKF model is usually described with insertion rates per link and deletion rates per link. In that parameterization the insertion and deletion rates can differ. We call the deletion rate scaled per nucleotide θ.

When computing a pairwise distance, the gaps-as-missing-data correction removes sites in which either sequence has a gap from consideration. The expected length of a locus under the paired invariants model will be denoted N. This will be a function of the expected length of each block of variable sites, which is a function of the insertion rate relative to the deletion rate. Our argument applies to any insertion rate which leads to a non-infinite equilibrium sequence length. So we phrase the argument in terms of the per-locus expected length and do not use the insertion rate parameter explicitly in our argument.

Under the TKF91 model, each block of variable sites is expected to follow a geometric distribution with a parameter that depends on the ratio of the per-link insertion and deletion rates. Because sites with an insertion and then a deletion are typically culled from an alignment, we consider a pairwise alignment length to be the length of the correct alignment after all positions with gaps in both members of the pair are removed. Even though the expected number of nucleotides in each sequence does not change, the insertion of new positions and deletion of sites means that the pairwise alignment length grows as a function of the evolutionary distance. In the paired invariants model, let $a(t, \theta, p_{\text{inv}})$ denote the expected length of a pairwise alignment of two sequences separated by path length, t. Then:

$$a(0, \theta, p_{\text{inv}}) = N \tag{1}$$

$$\lim_{t\to\infty} a(t, \theta, p_{\text{inv}}) = N(p_{\text{inv}} + 2(1 - p_{\text{inv}})) \tag{2}$$

where Np_{inv} is the number of invariable columns in the alignment. $N(1 - p_{\text{inv}})$ columns are expected to be in the ancestor but deleted along the path to the descendant. Because the process started at equilibrium, we expect them to be replaced by $N(1 - p_{\text{inv}})$ inserted sites.

For each site that is free to vary in the ancestor, the probability that it survives to the descendant is $e^{-\theta t}$, using the exponential distribution. We refer to columns where there is a nucleotide in both the ancestor and the descendant as "ungapped columns". The expected number of ungapped columns is

$$b(t, \theta, p_{\text{inv}}) = N\left(p_{\text{inv}} + (1 - p_{\text{inv}})e^{-\theta t}\right)$$

Note that $\lim_{t\to\infty} b(t, \theta, p_{\text{inv}}) = Np_{\text{inv}}$.

The expected proportion of *residues* in a sequence which are free to vary remains constant at $1 - p_{\text{inv}}$ as branch length approaches infinity. However, if we consider only ungapped columns in the true alignment of two sequences, we see that the proportion of these sites which are variable approaches 0 as deletions continue to reduce the number of aligned columns among the class of variable sites. The expected proportion of ungapped columns that are free to vary is:

$$\begin{aligned}\Pr(\text{variable}|\text{ungapped}) &= \frac{N(1 - p_{\text{inv}})e^{-\theta t}}{b(t, \theta, p_{\text{inv}})} \\ &= \frac{(1 - p_{\text{inv}})e^{-\theta t}}{p_{\text{inv}} + (1 - p_{\text{inv}})e^{-\theta t}}\end{aligned} \tag{3}$$

This function is plotted in Fig. 2A for the case when $p_{\text{inv}} = 0.2$ and $\theta = 0.1$.

Recall that under the Jukes–Cantor model the probability of a site having a different nucleotide from its ancestor across a path of length t is $\frac{3}{4}\left(1 - e^{-\frac{4}{3}t}\right)$. For the Jukes–Cantor model with invariant sites the probability of a difference, conditional on a site being a member of the variable class is:

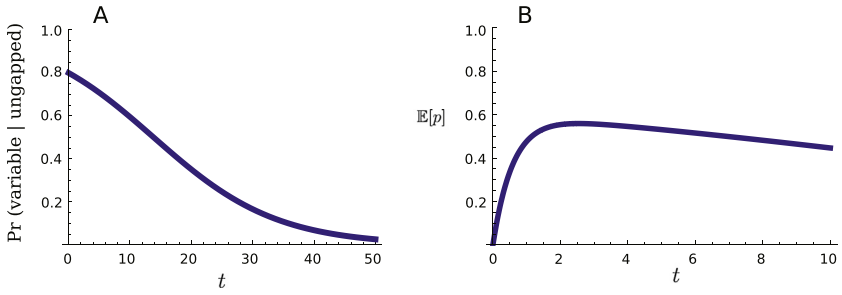

Fig. 2. Properties of the paired invariants model with $p_{inv} = 0.2$ and $\theta = 0.1$. A. The proportion of aligned sites which are free to vary as a function of time (Eq. (3)) B. Pairwise nucleotide substitution distance through time (Eq. (5)). Note 5-fold difference in t-axis scale between A and B.

$$\Pr(\text{difference}|\text{ungapped, variable}) = \frac{3}{4}\left(1 - e^{\frac{-4t}{3(1-p_{inv})}}\right). \quad (4)$$

The only difference between this formula and the Jukes–Cantor transition probability is the inclusion of a $1 - p_{inv}$ factor to increase the rate of substitution for the variable sites. This is included to adhere to the common convention that the mean rate of substitutions is equal to 1.0 per site.

For a pair of sequences, the probability of seeing a different state at a randomly chosen, ungapped, variable site (Eq. (4)) is a monotonically increasing function of t. However, the proportion of ungapped sites which are variable decreases, as was shown in Eqn. (3). The expected pairwise distance for the paired invariants model when measured as the expected proportion of ungapped positions that differ between the tips is:

$$\begin{aligned}\mathbb{E}[p] &= \Pr(\text{difference}|\text{ungapped, variable})\Pr(\text{variable}|\text{ungapped}) \\ &= \frac{3(1-p_{inv})e^{-\theta t}\left(1 - e^{\frac{-4t}{3(1-p_{inv})}}\right)}{4(p_{inv} + (1-p_{inv})e^{-\theta t})}.\end{aligned} \quad (5)$$

This expected p-distance is shown in Fig. 2B. Note that it is not a monotonically increasing function.

2.4. Gaps-as-missing distance correction

Under a gaps-as-missing analysis, only the ungapped columns are relevant in distance calculations. Thus, the expected p-distance shown in Eq. (5) fills the role of $g(t)$ in the discussion of our proofs about the consistency of distance-based tree estimation. Note that the substitution model for the paired invariant sites model is simply the Jukes–Cantor substitution model with invariant sites. If we assume that we know the (correct) proportion of invariant residues in the generating process, then the distance correction for this model is:

$$f^{-1}(p) = \frac{-3(1-p_{inv})}{4}\ln\left(1 - \frac{4p}{3(1-p_{inv})}\right). \quad (6)$$

We can combine Eqs. (5) and (6) to express the transformed evolutionary distances $\bar{t}$ as a function of the true evolutionary distance, t:

$$\begin{aligned}\bar{t} &= \frac{-3(1-p_{inv})}{4}\ln\left(1 - \frac{4\frac{(1-p_{inv})e^{-\theta t}\left(\frac{3}{4} - \frac{3}{4}e^{\frac{-4t}{3(1-p_{inv})}}\right)}{p_{inv} + (1-p_{inv})e^{-\theta t}}}{3(1-p_{inv})}\right) \\ &= \frac{-3(1-p_{inv})}{4}\ln\left(1 - \frac{e^{-\theta t}\left(1 - e^{\frac{-4t}{3(1-p_{inv})}}\right)}{p_{inv} + (1-p_{inv})e^{-\theta t}}\right)\end{aligned} \quad (7)$$

This function is shown in Fig. 3.

Clearly the function is not linear; indeed it is not monotonically increasing. In fact, the function is not linear even at small path lengths. The first and second derivatives of the distance correction with respect to t (see Appendix B) are somewhat complicated. However, when evaluated at $t = 0$, the first derivative is 1 and the second derivative of the expected value of the distance correction is $-2p_{inv}\theta$. Thus, the gaps-as-missing-data approach coupled with the correct substitution model results in a concave distance correction function whenever both $p_{inv} > 0$ and $\theta > 0$. Lemma 3 of Appendix A states that this will lead to statistically inconsistent estimation of the tree topology for some tree shapes.

The proof described above only applies to the four-tip tree used in the construction of the argument, but there is no reason to be confident that larger trees will not be susceptible to similar effects. To demonstrate this, we performed a simulation of a six-taxon tree that is constructed by replacing each tip at the end of a long branch in a Felsenstein-zone tree with a pair of taxa. We simulated 100 datasets on this tree shape under a JC + TKF91 paired-invariants model with an equilibrium sequence length of 100,000 and $p_{inv} = 0.5$ (see Appendix C for simulation details). Estimation of the trees under the weighted least-squares criterion using PAUP* v4.0b10 (Swofford, 2001) with the JC distance correction and an assumption that $p_{inv} = 0.5$ preferred the long-branch attraction tree topology in all 100 simulation replicates.

We implemented a simple program which estimates the equilibrium sequence length by taking an average of the (non-gap) length of each sequence. The software then removes a number of constant columns that is compatible with the assumption that $p_{inv} = 0.5$ in the paired-invariants model. These culled matrices serve as a reasonable proxy for the free-to-vary sites in the simulation. Weighted least-squares tree inference was performed on all of these culled data sets using the JC distance correction. The correct tree was recovered in 86 out of the 100 cases. This demonstrates that, in most cases, there was sufficient information to estimate the true tree correctly if one analyzes the data in a manner that is compatible with the paired invariants model.

We conjecture that this pipeline of

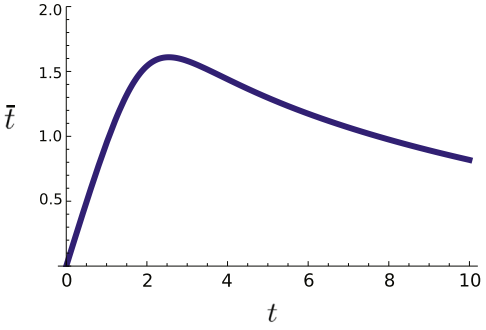

Fig. 3. The transformed evolutionary distance $\bar{t}$ values as a function of true evolutionary distance t, under the paired invariants model with $p_{inv} = 0.2$ and $\theta = 0.1$ (Eq. (7)).

1. estimating the equilibrium sequence length using the average sequence length, $\hat{N}_e$,
2. culling $p_{\text{inv}}\hat{N}_e$ constant, gapless characters (based on an assumed value of p_{inv}), and
3. applying a distance correction that uses the correct substitution model and no invariant sites

would provide a statistically consistent estimate of an additive distance matrix that would lead to statistically consistent estimate of the phylogeny. However, because this procedure requires knowledge of the correct value of p_{inv}, it does not represent a viable solution for real-world analyses.

3. Conclusions

We have extended the work of Susko et al. (2004) by proving that there is a tree shape which will lead to the positively misleading estimation of an incorrect tree topology when the distance correction function is convex. We have also proven that the commonly applied gaps-as-missing-data approach will not lead to statistical inconsistency of distance estimates if the indel and substitution processes are independent. However, sequence evolution follows the paired invariants model, the deviation from independence is sufficient to lead to inconsistency of the distance estimates and the tree topology.

Obviously, the paired invariants model with a Jukes–Cantor substitution process is an extremely simple model which does not accurately depict the evolution of real sequences. Nevertheless, the paired invariants model encapsulates a simple idea that has been at the core of thinking about molecular evolution ever since Kimura (1968): constant sites probably are constrained because they play an important functional role. It seems entirely plausible that the subset of functionally important sites in the genome would be prevented from experiencing fixation of indels or substitutions. Thus it is troubling that adding this idea to the simplest possible substitution model is sufficient to lead to inconsistency of phylogenetic inference.

One solution would be to rely on distance corrections which do not treat gaps as missing data. Another option may be using multiple values of p_{inv} to correct for the fact that the proportion of ungapped positions which correspond to constrained sites is likely to be higher for comparisons over long evolutionary timespans. Both the proportion of gapped sites in the correct pairwise sequence alignment and the proportion of ungapped positions which are variable (shown in Fig. 2A) are monotonically changing functions of the path length. This implies that it may be possible to devise some recipe for correcting distances that uses a pair-specific value of p_{inv}, and that this pair-specific p_{inv} could be calculated from an observable property of an alignment. Such a procedure might rescue distance-based from inconsistency when the data are generated by the paired invariants model. However, this form of inference would probably be sensitive to slight inadequacies of the model because accounting for rate heterogeneity when using pairwise data alone is notoriously difficult. Our results underscore that fact that phylogenetic inference is a problem that is so difficult that even subtle forms of ascertainment bias can lead to fundamental misbehavior of inference methods.

Acknowledgments

EJM was supported by an Alexander von Humboldt award. MS was supported the NZ Marsden Fund and the Allan Wilson Centre. MTH was supported by ATOL-0732920, the AVAToL Open Tree of Life award and HITS.

Appendix A

Suppose that distances are generated on a tree by a model M_g and corrected assuming a model M_f.

Theorem 1. *Suppose that $f(t)$ and $g(t)$ (the functions for correcting and generating p-distances respectively) are analytic functions of t that are strictly increasing in some neighborhood of 0, and satisfy $f(0) = g(0) = 0$. Let $h(t) = \bar{t} = f^{-1}(g(t))$ (the transformed evolutionary distances). Then precisely one of the following conditions holds:*

- *The correction process f is equal to the generating function g up to a scalar multiple (i.e. $f(t) = g(t/c)$ and so $h(t) = ct$ for all $t \in [0, \rho)$, for some constant c). In this case NJ will select the correct tree topology when applied to the transformed evolutionary distances; or*
- *The correction process f is not equal to the generating function g up to a scalar multiple. In this case there exists a binary tree on four leaves with an associated set of strictly positive branch lengths for which NJ will select an incorrect tree topology when applied to the transformed distances.*

The proof of this result involves combining five lemmas; the first is standard, the second is a formal statement of results from Susko et al. (2004), the third is new, and the fourth and fifth are technical lemmas.

Lemma 2 Saitou and Nei (1987, p. 413). *NJ applied to distance data on four taxa (A, B, C, D) returns the quartet tree $AB|CD$ if $d_{AB} + d_{CD} < \min\{d_{AC} + d_{BD}, d_{AD} + d_{BC}\}$.*

Lemma 3. *Suppose the transformed distance function $h(t)$ is strictly concave and increasing on the interval $[\lambda, 2\lambda]$ for some $\lambda > 0$. For any $\sigma > 0$ sufficiently small, distances on Felsenstein-Zone tree of Fig. 1(A) that are transformed by h have the property that NJ will estimate the incorrect tree topology (AD|BC).*

Proof of Lemma 3. By Lemma 2, for any distance function d on four taxa i, j, k, l, NJ applied to d will return the quartet tree $ij|kl$ when i, j minimizes the pairwise sum $d_{ij} + d_{kl}$. Let us now put $d_{ij} = h(t_{ij}) = \bar{t}_{ij}$ (i.e. the transformed evolutionary distances). Consider the three pairwise sums:

(S1) $d_{AB} + d_{CD} = 2h(\lambda + \sigma)$;
(S2) $d_{AC} + d_{BD} = 2h(\lambda + 2\sigma)$;
(S3) $d_{AD} + d_{BC} = h(3\sigma) + h(2\lambda + \sigma)$;

Since h is strictly increasing on $[\lambda, 2\lambda]$, the expression (S2) is always greater than (S1) for any $\sigma > 0$. Thus it suffices to show that case (S3), which corresponds to NJ returning the tree $AD|BC$, is less that (S1) for sufficiently small $\sigma > 0$. To this end, note that since h is strictly concave on $[\lambda, 2\lambda]$ we have: $h(2\lambda) < 2h(\lambda)$, so if we let

$$q(x) := 2h(\lambda + x) - h(2\lambda + x) - h(3x)$$

then $q(0) = 2h(\lambda) - h(2\lambda) - h(0) > 0$ (recall $h(0) = 0$). Since h is continuous (by virtue of being analytic) q is too, so it follows that for any sufficiently small (but strictly positive) value of σ we have $q(\sigma) > 0$. Because $q(\sigma)$ equals the quantity described by (S1) minus that described by (S3), when $q(\sigma) > 0$ then NJ will prefer tree $AD|BC$ over the true tree $AB|CD$. □

Lemma 4. *Suppose the transformed distance function $\bar{t} = h(t)$ is strictly convex and increasing on the interval $[\lambda, 2\lambda]$ for some $\lambda > 0$. For any $\sigma > 0$ sufficiently small, distances on the "twisted*

Farris-zone" tree of Fig. 1(B) that are transformed by h have the property that neighbor-joining will estimate the incorrect tree topology (AD|BC).

Proof of Lemma 4. For the "twisted Farris-zone" tree of Fig. 1(B) consider the three pairwise sums:

(S1) $d_{AB} + d_{CD} = h(2\lambda + \sigma) + h(3\sigma)$;
(S2) $d_{AC} + d_{BD} = h(\lambda + 4\sigma) + h(\lambda + 2\sigma)$;
(S3) $d_{AD} + d_{BC} = 2h(\lambda + 3\sigma)$.

Now, if h is strictly convex on $[\lambda, 2\lambda]$ and if $x, y \in [\lambda, 2\lambda]$ then:

$$h\left(\frac{x+y}{2}\right) < \frac{1}{2}[h(x) + h(y)].$$

Applying this with $x = \lambda + 4\sigma$ and $y = \lambda + 2\sigma$, where $0 < \sigma < \lambda/4$, gives:

$$h(\lambda + 3\sigma) < \frac{1}{2}[h(\lambda + 4\sigma) + h(\lambda + 2\sigma)],$$

which gives (S3)<(S2).

Again by convexity, $h(2\lambda) > 2h(\lambda)$, so if we let

$$q(x) := h(2\lambda + \sigma) + h(3\sigma) - 2h(\lambda + 3\sigma).$$

then $q(0) = h(2\lambda) - 2h(\lambda) + h(0) > 0$. By a similar continuity argument as in the concave case, $q(\sigma) > 0$ for all $\sigma > 0$ sufficient close to 0. Because $q(\sigma)$ equals the quantity described by (S1) minus that described by (S3), if we take $\sigma \in (0, \lambda/4)$ to be small enough that $q(\sigma) > 0$ (i.e. (S3)<S(1)), and recall from above that for $0 < \sigma < \lambda/4$ we also have (S3)<(S2), then this provides conditions for which NJ will again prefer tree $AD|BC$ over the true tree $AB|CD$. □

Lemma 5. *Under the assumptions on f and g in Theorem 1, the transformed distance function h is a strictly increasing analytic function of t on $[0, \rho)$ for some $\rho > 0$.*

Proof of Lemma 5. The proof that h is analytic is straightforward, since analytic functions (in particular f and g) are closed under composition, and also under functional inverse (providing their derivative is non-zero, as it is here). To see that h is increasing, at least close to 0, note that, by elementary differential calculus, we have:

$$\frac{d}{dt}h(t) = \frac{g'(t)}{f'(f^{-1}(g(t)))}. \tag{8}$$

By assumption, f and g are both increasing in some neighborhood of 0, and since $f^{-1}(g(0)) = f^{-1}(0) = 0$, there exists $\rho > 0$ for which the numerator and denominator of Inequality (8) are both strictly positive for all $t \in [0, \rho)$. □

Lemma 6. *Suppose $H(t)$ is a real-valued function that is analytic in $[0, \rho)$ for some $\rho > 0$, and that satisfies $H(0) = 0$. If $H(t) \neq ct$ on $[0, \rho)$ for some constant c, then there exists some value $s > 0$ so that $H(t)$ is either strictly concave on the interval $[s/2, s]$ or strictly convex on the interval $[s/2, s]$.*

Proof of Lemma 6. If $H''(0) > 0$ then since H'' is continuous at 0, there is a value $s \in [0, \rho)$ so that $H''(t) > 0$ for all $t \in [0, s]$ and so H is strictly convex on $[s/2, s]$. Similarly, if $H''(0) < 0$ then H is strictly concave on $[s/2, s]$ for some $s > 0$. Suppose that $H''(t) = 0$. Then either (i) there exists a smallest $k > 2$ for which $H^{(k)}(0) \neq 0$ (call this value k_1) or (ii) $H^k(0) = 0$ for all $k > 2$. In Case (i), suppose first that $a := H^{(k_1)}(0) > 0$. A Taylor series expansion of H about 0 gives $H(t) = at^{k_1} + \cdots$ where the remaining terms are of order t^{k+1} and higher. Thus, for a sufficiently small $v \in (0, \rho)$, we have $H''(t) = k_1(k_1 - 1)at^{k_1 - 2} +$ (terms of order $t^{k_1 - 1}$ and higher) so $H''(t) > 0$ for all $t \in (0, v)$. In particular, for any strictly positive value of s less than v we have $H''(t) > 0$ for all $t \in [s/2, s]$. Thus, as before, H'' is strictly convex on $[s/2, s]$. A similar argument (for strict concavity) applies if $H^{(k_1)}(0) < 0$. In Case (ii) the Taylor expansion of $H(t)$ on $[0, \rho)$ centered on 0, shows that $H''(t) = 0$ for all $t \in [0, \rho)$. By integrating (twice) it follows that $H(t) = ct + H(0)$ for all $t \in [0, \rho)$, for some constant c. Since $H(0) = 0$, this gives $H(t) = ct$, as claimed. □

Proof of Theorem 1. By Lemma 5, h and analytic and increasing in $[0, \rho)$, so by Lemma 6, if we take $H(t) = h(t)$ then if h is not linear, it is either strictly concave or strictly convex on an interval of the form $[s/2, s]$ for some $s \in (0, \rho)$. Theorem 1 now follows from Lemmas 3 and 4. □

Appendix B

The first and second derivatives of the expected corrected distance (Eq. (7)) with respect to the path length t are:

$$c[t] = e^{\frac{4t}{3-3p_{\text{inv}}}}$$

$$d[t] = e^{t\left(\theta + \frac{4}{3-3p_{\text{inv}}}\right)}$$

$$\frac{\partial \bar{t}}{\partial t} = \frac{4 - 3(e^{\theta t} - d[t])p_{\text{inv}}^2\theta + p_{\text{inv}}(e^{\theta t}[4+3\theta] - 4 - 3d[t]\theta)}{4(1 - c[t]p_{\text{inv}} + d[t]p_{\text{inv}})(1 + e^{\theta t}p_{\text{inv}} - p_{\text{inv}})}$$

$$v[t] = e^{t\left(\theta + \frac{8}{3-3p_{\text{inv}}}\right)}$$

$$w[t] = e^{t\left(3\theta + \frac{8}{3-3p_{\text{inv}}}\right)}$$

$$x[t] = e^{t\left(3\theta + \frac{4}{3-3p_{\text{inv}}}\right)}$$

$$y[t] = e^{2t\left(\theta + \frac{2}{3-3p_{\text{inv}}}\right)}$$

$$m = p_{\text{inv}} - 1$$

$$\begin{aligned} u[t] = &-16c[t]m^2 + 9e^{\theta t}m^3\theta^2 + 9v[t]m^3p_{\text{inv}}\theta^2 - 9w[t]m^2p_{\text{inv}}^2\theta^2 \\ &+ x[t]p_{\text{inv}}^2(4 - 3m\theta)^2 + 16y[t]p_{\text{inv}}(2 + 3\theta + 3p_{\text{inv}}^2\theta - 3p_{\text{inv}}[1 + 2\theta]) \\ &- d[t]m\left(3p_{\text{inv}}^2(8 - 3\theta)\theta + 9p_{\text{inv}}^3\theta^2 + (4 + 3\theta)^2 - 3p_{\text{inv}}[16 + 16\theta + 3\theta^2]\right) \end{aligned}$$

$$\frac{\partial^2 \bar{t}}{\partial t^2} = \frac{p_{\text{inv}}u[t]}{12m(1 - c[t]p_{\text{inv}} + d[t]p_{\text{inv}})^2(1 + e^{\theta t}p_{\text{inv}} - p_{\text{inv}})^2}$$

These were calculated using the Mathematica notebook included as part of the supplementary materials.

Appendix C

Sequences were simulated on a Felsenstein-zone shaped tree with a newick representation:

```
((A:0.025,A1:0.025):0.375,B:0.025,((C:0.025,C1:0.025)
:0.375,D:0.025):0.025);
```

The simulation model was the paired invariants model with the JC substitution process and $p_{\text{inv}} = 0.5$. An equilibrium sequence length of 100,000 was used. An insertion rate of 0.5 and a deletion rate of 1.0 were used for the variable blocks. This leads to a geometric

distribution of sequence lengths for the variable blocks with a mean length of 1.

Each dataset was analyzed using PAUP*'s commands:

```
dset dist = jc pinv = 0.5 objective = lsfit;
alltrees;
```

to conduct an exhaustive search of tree using the weighted least-squares criterion and distances corrected under the correct substitution model and value of p_{inv}.

To demonstrate that there was sufficient information in the simulated data to identify the tree, we implemented a method of culling a proportion of the constant, gapless sites from a matrix by assuming that the data were generated under the paired invariants model with a particular value of p_{inv}. This script is `paired_invariants_cull.py` of the DendroBites repository hosted at https://github.com/mtholder/DendroBites.

Culled datasets were analyzed in PAUP* using the commands:

```
dset dist = jc pinv = 0 objective = lsfit;
alltrees;
```

to conduct an exhaustive search using the weighted least-squares criterion and the JC distance correction with no among-site rate heterogeneity.

Appendix D. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.07.027.

References

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27, 401–410.

Grievink, L.S., Penny, D., Holland, B.R., 2013. Missing data and influential sites: choice of sites for phylogenetic analysis can be as important as taxon sampling and model choice. Genome Biol. Evol. 5, 681–687.

Jukes, T.H., Cantor, C.R., 1969. In: Evolution of Protein Molecules. Elsevier, pp. 21–132.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.

Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214.

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

Soubrier, J., Steel, M., Lee, M., Der Sarkissian, C., Guindon, S., Ho, S., Cooper, A.the Genographic Consortium, 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. 29, 3345–3358.

Susko, E., Inagaki, Y., Roger, A.J., 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol. Biol. Evol. 21, 1629–1642.

Swofford, D.L., 2001. PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) 4.0.b5. Sunderland, Massachusetts: Sinauer Associates.

Tajima, F., 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 10, 677–688.

Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33, 114–124.

Warnow, T., 2012. Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. PLoS Currents 4.