Molecular Phylogenetics and Evolution xxx (2015) xxx–xxx

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 xxxx

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: Heidelberg Institute for Theoretical Studies, Heidelberg, Germany.

E-mail address: ejmctavish@gmail.com (E.J. McTavish).

http://dx.doi.org/10.1016/j.ympev.2015.07.027
1055-7903/© 2015 Elsevier Inc. All rights reserved.

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.

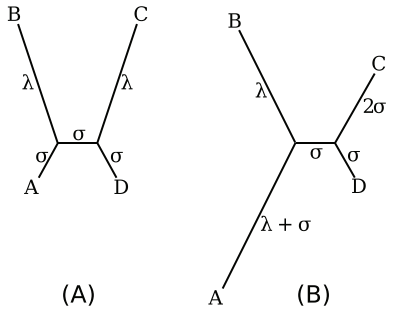

Fig. 1. (A) The Felsenstein-zone tree with branch lengths used in the proof of Lemma 3; (B) The "twisted Farris-zone" tree used in the proof of Lemma 4.

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 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