On Inconsistency of the Neighbor-Joining, Least Squares, and Minimum Evolution Estimation When Substitution Processes Are Incorrectly Modeled

Edward Susko,* Yuji Inagaki,† and Andrew J. Roger†

*Genome Atlantic, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada; †Genome Atlantic, Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Using analytical methods, we show that under a variety of model misspecifications, Neighbor-Joining, minimum evolution, and least squares estimation procedures are statistically inconsistent. Failure to correctly account for differing rates-across-sites processes, failure to correctly model rate matrix parameters, and failure to adjust for parallel rates-across-sites changes (a rates-across-subtrees process) are all shown to lead to a “long branch attraction” form of inconsistency. In addition, failure to account for rates-across-sites processes is also shown to result in underestimation of evolutionary distances for a wide variety of substitution models, generalizing an earlier analytical result for the Jukes-Cantor model reported in Golding (Mol. Biol. Evol. 1:125–142, 1983) and a similar bias result for the GTR or REV model in Kelly and Rice (1996). Although standard rates-across-sites models can be employed in many of these cases to restore consistency, current models cannot account for other kinds of misspecification. We examine an idealized but biologically relevant case, where parallel changes in rates at sites across subtrees is shown to give rise to inconsistency. This changing rates-across-subtrees type model misspecification cannot be adjusted for with conventional methods or without carefully considering the rate variation in the larger tree. The results are presented for four-taxon trees, but the expectation is that they have implications for larger trees as well. To illustrate this, a simulated 42-taxon example is given in which the microsporidia, an enigmatic group of eukaryotes, are incorrectly placed at the archaea-bacteria-eukaryotes split because of incorrectly specified pairwise distances. The analytical nature of the results lend insight into the reasons that long branch attraction tends to be a common form of inconsistency and reasons that other forms of inconsistency like “long branches repel” can arise in some settings. In many of the cases of inconsistency presented, a particular incorrect topology is estimated with probability converging to one, the implication being that measures of uncertainty like bootstrap support will be unable to detect that there is a problem with the estimation. The focus is on distance methods, but previous simulation results suggest that the zones of inconsistency for distance methods contain the zones of inconsistency for maximum likelihood methods as well.

Introduction

Distance methods for the estimation of phylogenetic trees such as the Neighbor-Joining method (NJ; Saitou and Nei 1987), and least-squares estimation (LS; Cavalli-Sforza and Edwards 1967) are valuable tools because of their computational simplicity. In large-scale problems, or when repeated fitting is required, maximum likelihood (ML) estimation can become infeasible, with distance methods providing a reasonable alternative. Such cases arise when competing models of molecular evolution are being considered, when large numbers of sequences are under consideration, when bootstrapping is required and in simulation studies. Because of the importance of distance methods, investigations of situations where these methods fail to consistently estimate topologies is valuable. In the absence of model misspecification most distance methods are consistent. Consistency of distance methods has been proven by Saitou and Nei (1987) for NJ, and by Rzhetsky and Nei (1993) for minimum evolution (ME) and LS. However some distance methods like UPGMA and ME with weighted least-squares branch lengths can be inconsistent even in the absence of model misspecification (Gascuel, Bryant, and Denis 2001).

We consider here whether the topology will be consistently estimated when an incorrect model or no model has been used in constructing the pairwise distances. Sequence evolution in nature is an extremely complex and heterogeneous (across sites and across lineages) process. In most cases, therefore, when inferring phylogeny from sequences, models with simplifying assumptions regarding rate matrices, rates-across-sites, or stationarity of the rates-across-sites process across lineages will always be incorrect to some degree. For this reason, it is of critical importance to investigate the performance of methods when their models are misspecified in various ways. Inconsistency of a method of phylogenetic estimation implies poor performance as the number of sites gets large, and as such it is at least theoretically possible that a method that is inconsistent will perform reasonably with small numbers of sites. This is rarely the case in practice; inconsistency is usually a good indicator of situations in which poor performance can be expected, even with small numbers of sites. Indeed, for the forms of model misspecification similar to those considered through simulation with finite samples by Gaut and Lewis (1995) and Huelsenbeck (1995), the zones of inconsistency match up well with regions of poor performance by the methods considered by those authors.

The results presented indicate that serious difficulties of inference often arise in the presence of inconsistency. In many of the cases considered we show that with probability converging to one, a particular incorrect topology is estimated. As a consequence, with larger numbers of sites, there will be large bootstrap support for this incorrect topology. To illustrate briefly, we simulated data from the tree \((A : 0.8, C : 0.1) : 0.1, (B : 0.8, D : 0.1)\) with a Jukes-Cantor (JC; Jukes and Cantor 1969) substitution process. As indicated later, with
probability one, NJ with Hamming or \( p \)-distance will estimate the long branch topology, characterized by the neighbors \((A, B)\), with a large enough number of sites. Even with 200 sites, the probability is very large that the long branch topology will be estimated. In 1,000 simulations the long branch topology \((A, B)\) was estimated every time. More significantly, the average bootstrap support for the topology \((A, B)\) over the 1,000 simulations was 81.3%, so that there was little information to suggest that there was a problem.

The analytical nature of the results presented lend insight into the reasons for inconsistency. The form of inconsistency most commonly discussed in the literature is long branch attraction in the context of the maximum parsimony method (Felsenstein 1978; Hendy and Penny 1989). Long branch attraction turns out to be the main form of inconsistency that we find occurring as well. The reason for this is that in so many of the cases where an overly simple model or no model is used to construct the distances, the limiting distances turn out to be a concave function of the actual distances. In contrast, when the limiting distances are convex functions of the actual distances, long branch inconsistency will not arise [a continuous concave (convex) function has negative (positive) second derivative everywhere or roughly speaking, has an upside-down (right-side up) bowl shape]. Instead, as we illustrate with an example, the form of inconsistency that can arise in this case is a “long branches repel” type of inconsistency.

Because it is analytically tractable and eases discussion, we primarily consider four taxon topologies. However a form of parallel rates-across-sites variation model misspecification is considered both analytically and, through simulation, with more taxa. Although idealized, the basic notion of parallel rates-across-sites variation is biologically relevant as a model for the evolutionary implications of differing functional constraints for different organisms. For instance, Steel, Huson, and Lockhart (2000) discussed a case in which changes in the numbers of variable sites in 18S ribosomal DNA from groups of Holometabolus insects provided artifactual support for Strepsiptera as a sister taxon of Diptera. In this study, we consider a 42-taxon example where a group of intracellular parasitic eukaryotes, the microsporidia, are erroneously considered a 42-taxon example where a group of intracellular parasitic eukaryotes, the microsporidia, are erroneously

Limiting Distances and Assumptions

Consider a pair of taxa, \( i \) and \( j \). Distances between the two taxa are computed from the relative frequencies \( F_{ij}^{(ij)} \), with which character states \( x \) and \( y \) arise in the two taxa across all sites. In some cases additional parameters might be required to compute the distance between taxa. For instance, computing gamma-corrected distances requires an estimate of the \( \alpha \) parameter (c.f. Nei, Chakraborty, and Rice, 1996) and confirming the simulation results of Gaut and Rice (1996) and confirming the simulation results of Gascuel (1994) and Huelsenbeck (1995), Kuhner and Felsenstein (1994), and Bruno and Halpern (1999). On the basis of this and other work, it appears that there is widespread understanding that model misspecification can lead to biased estimation and inconsistency. Theoretical work that explicitly considers consistency in the case of Hamming distances includes reports by DeBry (1992) and Rzhetsky and Sitnikova (1996). Models where rates vary across sites and subtrees have been considered theoretically by Chang (1996) who shows that inconsistency can arise as branch length converges to zero. The work here can be considered an addition to the present work on the topic that (1) lends analytical insight into the reasons for the long branch form attraction that so frequently arises, (2) provides illustrations of alternative forms of model misspecification that can lead to inconsistency (specifically parallel rates-across-sites variation) that are not easily adjusted for, and (3) more closely considers the possible inconsistency of NJ as well as its relationship to LS and ME.

An outline of the article is as follows. The next two sections are technical, introducing assumptions, notation, and the main derivations leading to a zone of inconsistency. The following two subsections give examples of zones of inconsistency arising from a failure to model rates-across-sites and rates-across-subtrees processes. We also show in these sections that failure to account for rates-across-sites variation results in underestimation of evolutionary distances for a variety of models generalizing previous results for the JC model (Golding 1983), GTR model (Kelly and Rice, 1996) and confirming the simulation results of Kuhner and Felsenstein (1994). Other examples are then given of inconsistency corresponding to misspecifications in conventional model assumptions and resulting in long branches repelling. We conclude with a discussion of the relationships between the zones of inconsistency for NJ and those of LS and ME methods.

Downloaded from https://academic.oup.com/mbe/article-abstract/21/9/1629/992076 by guest on 12 April 2019
between taxa \(i\) and \(j\). As an example, Hamming distance or \(p\)-distance defined as the proportion of sites whether the character states for the two taxa differ, depends only on the relative frequencies \(F_{ij}^{(x)}\) through the proportion of different character states. Because no additional parameters are required for the computation of Hamming distance, \(\theta\) and \(\Theta\) can be taken to equal 1. Here \(d_j(F(i,j), \Theta) = 1 - \sum F_{ij}^{(x)}\).

We assume generally that \(d_j(F(i,j), \Theta)\) is a continuous function of \(F^{(x)}\) and \(\Theta\). Let \(P^{(x)}\) be the probability that character states \(x\) and \(y\) arise in the two taxa across all sites. Since \(F^{(x)}\) converges to \(P^{(x)}\) as the number of sites goes to infinity, and by assumption \(\Theta\) converges to \(\theta\) we get that

\[
d_{ij}(F(i,j), \Theta) \to d_{ij}(P^{(ij)}, \theta)
\]
as the number of sites goes to infinity.

It will be valuable to think of incorrectly specified distances in terms of their dependence on the correctly specified distances. To do this, we must make the additional assumption that the distance between two taxa is a monotonically increasing transformation \(g(t)\) of the correctly specified distance or branch length. This assumption is equivalent to the statement that larger branch lengths should be associated with larger distances, which should always be the case for any reasonable distance. As an example, for the JC model, the Hamming distance along a branch is related to branch length through the transformation \(g(t) = 3/4 - 3 \exp(-4t/\theta)/4\). In most of the examples that we consider, \(g(t)\) is a concave transformation of \(t\), and so, unless stated otherwise, this is assumed as well. We will assume throughout most of the discussion that the true tree is \(((A : b, C : a) : a, (B : b, D : a))\):

![Diagram](https://academic.oup.com/mbe/article-abstract/21/9/1629/992076 by guest on 12 April 2019)

In this case the limiting distances, \(d_j(P^{(ij)}, \theta)\), as the number of sites goes to infinity, are as given in table 1.

### Zones of Inconsistency: Neighbor Joining

In the case of a four-taxon tree with taxa \(A, B, C,\) and \(D\), there are three topologies which can be described in terms of the neighbor of \(A\): \((A, B), (A, C),\) and \((A, D)\).

<table>
<thead>
<tr>
<th>Distance</th>
<th>Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d_{AB}(P, \theta))</td>
<td>(g(a + 2b))</td>
</tr>
<tr>
<td>(d_{AC}(P, 0))</td>
<td>(g(2a + b))</td>
</tr>
<tr>
<td>(d_{AD}(P, 0))</td>
<td>(g(a))</td>
</tr>
<tr>
<td>(d_{BC}(P, 0))</td>
<td>(g(2a + b))</td>
</tr>
<tr>
<td>(d_{BD}(P, 0))</td>
<td>(g(a + b))</td>
</tr>
<tr>
<td>(d_{CD}(P, 0))</td>
<td>(g(3a))</td>
</tr>
</tbody>
</table>

Let \(d_{ij}\) be the distance between taxa \(i\) and \(j\). For a four-taxon tree the Neighbor-Joining algorithm can be shown (Saitou and Nei 1987) to be determined by the following rules:

1. \((A, C)\) is preferred to \((A, D)\) if
   \[
d_{AD} + d_{CB} - d_{BD} - d_{AC} > 0. \quad (1)
   \]

2. \((A, C)\) is preferred to \((A, B)\) if
   \[
d_{AB} + d_{CD} - d_{BD} - d_{AC} > 0. \quad (2)
   \]

Substituting the limiting distances from table 1 into (1) and (2) we obtain that, with probability 1, as the number of sites gets large, \((A, C)\) is preferred to \((A, D)\) if

\[
2[g(2a + b) - g(a + b)] > 0, \quad (3)
\]

and \((A, C)\) is preferred to \((A, B)\) if

\[
h(b;a) = g(2b + a) - 2g(a + b) + g(3a) > 0, \quad (4)
\]

where it is convenient to think of \(h(b;a)\) as a function of \(b\) for fixed \(a\). Since \(g\) is an increasing function, (3) is always satisfied so that, in the limit, the estimated topology is either \((A, C)\), or \((A, B)\), according to whether \(h(b;a) < 0\) or \(h(b;a) > 0\). For a concave function, \(g\), the derivative of \(g\) is a decreasing function so that

\[
h'(b;a) = 2[g'(2b + a) - g'(a + b)] < 0.
\]

Thus \(h(b;a)\) is a decreasing function of \(b\) and \(b\) such that \(h(b;a) < 0\) whenever \(b > b_0(a)\) in which case \(h(b;a) > 0\) for all \(b < b_0(a)\) and \(h(b;a) < 0\) for \(b > b_0(a)\). The zone of inconsistency can then be found by determining the solution, \(b_0(a)\), of the equation \(h(b;a) = 0\); we set \(b_0(a) = \infty\) if no solution exists. For fixed \(a\) the topologies estimated by NJ in the limit are as follows:

![Diagram](https://academic.oup.com/mbe/article-abstract/21/9/1629/992076 by guest on 12 April 2019)
Fig. 1.—Plots of the zones of inconsistency when the generating model is a JC model and Hamming distances are used in the NJ estimation. The generating tree is \((A : b, C : a) : a, (B : b, D : a)\). For any \((a, b)\) above the curve, NJ inconsistently estimates the topology as \((A, B)\) when the generating tree is \((A : b, C : a) : a, (B : b, D : a)\). For points below the curve, NJ is consistent. Figure 1B gives the zone of inconsistency after transforming branch lengths to expected proportions of different character states in the two taxa at the ends of the branch. Additional axes have been added to figure 1B indicating the boundary in terms of divergence times. These divergence times were based on a rate of \(5 \times 10^{-9}\) substitutions per site per year.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Estimated Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b &lt; b_0(a))</td>
<td>((A, C))</td>
</tr>
<tr>
<td>(b &gt; b_0(a))</td>
<td>((A, B))</td>
</tr>
</tbody>
</table>

On the boundary, when \(b = b_0(a)\), regardless of the number of sites, both of the topologies \((A, C)\) or \((A, B)\) will have positive probability of being estimated.

Although in most cases \(b_0(a)\) must be determined numerically, for some distances an explicit solution for \(b_0(a)\) can be calculated. For instance, if the true model is a JC model but Hamming distance is used, then \(b_0(a)\) is determined as

\[
b_0(a) = \frac{-3 \log \left[ 1 - \sqrt{1 - \exp(-8a/3)} \right]}{4}.
\]

A plot of \(b_0\) as a curve depending on \(a\) is given in figure 1A. In this case the generating model was the JC model but the distance used was the Hamming distance. Any \((a, b)\) point above this curve corresponds to a tree \((A : b, C : a) : a, (B : b, D : a)\) for which NJ will incorrectly estimate the long branch tree \((A, B)\), and any point below this curve corresponds to a tree for which NJ is consistent.

Following Huelsenbeck (1995), we plot zones of inconsistency after transforming branch lengths to probabilities or expected long run proportions of different character states for the two taxa. We do this so that plots can more directly be related to sequence data for two taxa. In models of molecular evolution, branch lengths can be interpreted as the expected number of substitutions along the branch at a site. Although one cannot directly observe numbers of substitutions in two sequences, once can observe the proportions of different character states. Figure 1B gives the zone of inconsistency with this transformation. In figure 1B and figure 7 we also indicate divergence times. To make the conversion we use a rate of evolution equal to \(5 \times 10^{-9}\) substitutions per site per year (Li, 1997 p. 75; Page and Holmes 1997, p. 239; Penny et al. 2001). It should be noted that this rate of molecular evolution is an average over noncoding sequences in vertebrate genomes; for nonsynonymous changes in coding DNA, the scale would be multiplied by a factor ranging from 2 to greater than 100. In any case, such rates are unlikely to remain constant over the large time scale implied by the longest branch lengths we consider. Thus divergence times are shown for illustrative purposes only and should not be taken literally.

Effects of Ignoring Rates-Across-Sites Variation

Many phylogenetic models for sequence data assume independent Markov substitution processes across sites. Most of the parameters for these Markov processes are assumed constant across sites, but it has long been recognized that it is unreasonable to assume that the overall rate of evolution is constant across sites (Fitch and Markowitz 1970; Uzzel and Corbin 1971; Nei 1987). It is usually possible to adjust for the heterogeneity of rates across sites if one assumes a probability distribution for these rates (Gu and Li 1996; Waddell and Steel 1997). The corrections under a JC model are given in table 2 for some of the more common rates-across-sites (RAS) models. The zones of inconsistency for several RAS models when the distance is calculated without adjustment (assuming a single fixed rate) are given in figure 2A through 2C. In each case, the size of the zone of inconsistency increases with the magnitude of model misspecification.

One of the other consequences of failing to make RAS adjustments is that distances will be underestimated.
Golding (1983) shows that this is true for the JC model. Kelly and Rice (1996) show that branch lengths are biased downwards for the GTR model when stationary frequencies are not estimated. In fact, it is more generally true, holding for the following commonly used models: JC, K2P (Kimura 1980), F81 (Felsenstein 1981), F84 (Felsenstein 1984), and the GTR or REV model (Yang 1994). In the remainder of this subsection we sketch the reasons for this phenomenon. Full details are available in the Supplementary Material online.

To establish notation, we fix the pair of taxa $i$ and $j$ under consideration and leave as implicit the dependence of various quantities on $i$ and $j$. Let $t$ denote the true distance between the pair, $\Pi$ the diagonal matrix with the stationary probabilities of the character states along its diagonal, and $Q$ the rate matrix for the model under consideration. The number of additional parameters in $Q$ varies depending on the model under consideration. For instance, for the JC model there are no additional parameters and for the F84 model there is a transition-transversion ratio. The matrix of pairwise probabilities, with $xy$th entry the probability of observing character states $x$ and $y$ for the two taxa, is then $E[e^{Qrt}]$, where expectation is with respect to the random rate $r$. The estimate $\hat{t}$ of $t$ can be constructed from the observed frequencies or using ML estimation, but we assume that the estimated distance $\hat{t}$ and the rest of the parameters in

<table>
<thead>
<tr>
<th>Model</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>$0.75\alpha(1 - \rho(0.75)^{-1/\alpha} - 1)$</td>
</tr>
<tr>
<td>Gamma+I</td>
<td>$0.75\alpha(1 - \rho(1 - \pi)^{-1/\alpha} - 1)$</td>
</tr>
<tr>
<td>Fixed+I</td>
<td>$-0.75(1 - \pi) \log[1 - \rho(1 - \pi)^{-1/\alpha}]$</td>
</tr>
<tr>
<td>Fixed</td>
<td>$-0.75 \log[1 - \rho(0.75)]$</td>
</tr>
</tbody>
</table>

Table 2
The Distances Between Taxa for Several Common RAS Models Assuming a JC Substitution Process as a Function of $\rho$, the Proportion of Sites Where Character States Are Different, for the Two Taxa

![Graph A: Gamma model](image1)
![Graph B: Gamma + Invariant model](image2)
![Graph C: Fixed-rate + Invariant model](image3)

**FIG. 2.**—Plots of the zones of inconsistency when the generating model is a JC model with a rates-across-sites (RAS) process but JC distances are used without a RAS correction. Points above the curve correspond to trees where NJ is inconsistent. In figure 2A, the generating RAS model is a gamma model with parameter $\alpha$. In figure 2B, the generating RAS model is a gamma + invariable sites model; the $\alpha$ parameter is 1 for the gamma part of the model, and $p_0$ is the probability that a randomly selected site is invariable. In figure 2C, each site is either invariable or has a fixed rate in the generating RAS model; $p_0$ is the probability a randomly selected site is invariable.
the estimated rate matrix $Q$ are obtained through ML estimation without assuming a RAS model.

Now, the estimated distance $i$ always satisfies that

$$i = \frac{1}{\sum r_{ij} = 1}$$

The term $\hat{Q} = \hat{E}\log(\hat{Q}_a^{0i})$ is the estimate of the pairwise probability matrix. As is shown in the section titled Convergence of Pairwise Probability Estimates When Rates-Across-Sites Variation Is Ignored in the Supplementary Material Online, for the models listed above, this matrix converges to the actual pairwise probability matrix $\hat{E}(\log(\hat{Q}_a^{0i}))$. In addition, $\Pi$ converges to $\hat{\Pi}$ and so $i$ converges to

$$i_0 = \frac{1}{\sum r_{ij} = 1}$$

It follows from the argument given in Section 3.1 of Kelly and Rice (1996) that each of the entries $0 > \sum r_{ij} = 1$ is the proportion of nucleotides that are different. Thus

$$i_0 < \frac{1}{\sum r_{ij} = 1}$$

Interestingly we are unable to apply the argument above to the HKY85 model (Hasegawa, Kishino, and Yano 1985) and amino acid models like the PAM model (Dayhoff and Eck 1968; Dayhoff, Schwartz, and Orcutt 1979), where the matrix $Q$ is fixed. Although this suggests that the phenomenon of underestimation may not hold for these models in certain settings, most of our practical and simulation experience suggests that it does.

**Parallel Rates-Across-Sites Variation**

Just as failing to adjust for rates-across-sites processes can cause inconsistency, so can a failure to account for rate variation across subtrees. We consider here the zones of inconsistency that can arise if parallel rates-across-sites changes occur in subtrees of the larger tree. In the usual rates-across-sites model, a site must remain in the same rate class over the whole tree, whereas in the model considered here, rates vary across sites and across subtrees. By parallel rates-across-sites variation we refer to variation across sites where there is an increase in the probability that a given rate shift will occur in parallel in two subtrees. In the case of homogeneous rates-across-sites processes, gamma-corrected distances can be used to avoid inconsistency. For the parallel rates-across-sites variation considered here, however, such a correction will not in general work, because the rates-across-sites process varies with the pairs of taxa considered. Although the model presented here is a simple one, the notion of parallel rates-across-sites variation is biologically relevant as a model for the evolutionary implications of differing functional constraints for different organisms (see Lockhart et al. 1998; Inagaki et al. 2003). In the 42-taxon EF-1$\alpha$ example considered later, it seems possible that some sites may be constrained in eukaryotes but not in archaeabacteria or Microsporidia. In terms of distributional properties, this is suggestive of a certain frequency of larger rates-across-sites for archaeabacteria and microsporidia and an increased probability that these rate shifts will occur in parallel in the two subtrees.

We start by considering a four-taxon tree. The true tree is assumed to be of the form $(A : b, C : a) : a, (B : b, D : c : a)$. Rates are assumed constant along the middle branch and the branches leading to $C$ and $D$. The rate variation is along the branches leading to $A$ and $B$, and the rates-across-sites assigned to these branches are assumed to be positively correlated. Specifically, we assume that the rates at any given site are assigned to the branches according to the probabilities of table 3.

The entries are chosen so that the expected rate along each branch is 1 and so that the same marginal rate distribution applies to each branch. For this to be a valid distribution $\pi_{00} \leq \pi$, where $\pi_{00}$ is the probability that, at a site, 0 rates are assigned to both branches and, for a given branch, $\pi$ is the marginal probability that, at a site, a 0 rate is assigned to it. For the rates to be positively correlated, it is necessary to have $\pi_{00} \geq \pi^2$. Note that while the model has been specified in terms of the zero rates, the implication is that the branches leading to $A$ and $B$ will have certain probability at a site of having larger rates, 1/(1 $\pi$), in parallel, than the rate 1 at all of the other branches. We assume a JC substitution process and that the estimates of the pairwise branches are the uncorrected JC branch length estimates,

$$d_{ij} = -0.75 \log[1 - p_{ij}^{0.75}]$$

where $p_{ij}$ is the proportion of nucleotides that are different. Let $p(i, j)$ denote the path from $i$ to $j$, and let $r_b$, $l_b$ be the rates and branch lengths for branch $b$. Note that for the branches leading to $C$ and $D$, as well as the middle branch, these rates would be 1 and the branch lengths would be $a$, whereas for the other two branches the rates are random variables. As the number of sites gets large this proportion approaches the probability that two nucleotides are different, which can be shown to be

$$0.75 \left\{ 1 - E \left[ \exp \left( -4 \sum_{b \in p(i, j)} r_b l_b / 3 \right) \right] \right\},$$

the expectation here being with respect to the random rates. Substituting this into the JC distances, the limiting distances are
\[ d_{ij} \to -0.75 \log E \left[ \exp \left\{ -4 \sum_{b \in p(i,j)} r_b h_b/3 \right\} \right]. \]

Considering each distance separately, we get

\[
\begin{array}{|c|c|}
\hline
\text{Distance} & \text{Limiting Value} \\
\hline
d_{AB} & a - 0.75 \log E[e^{-4r_{AB}/3}] \\
\hline
d_{AD} & 2a - 0.75 \log E[e^{-4r_{AD}/3}] \\
\hline
d_{AC} & 2a - 0.75 \log E[e^{-4r_{AC}/3}] \\
\hline
d_{BC} & a - 0.75 \log E[e^{-4r_{BC}/3}] \\
\hline
d_{BD} & a - 0.75 \log E[e^{-4r_{BD}/3}] \\
\hline
\end{array}
\]

Here

\[ E[e^{-4b}] = \pi + (1-\pi)\exp[-4b/(1-\pi)] \] (6)

and

\[ E[e^{-4r_{AB}+r_{AD}}] = \pi_{00} + 2(\pi - \pi_{00})e^{-4(1-\pi)b/3} \\
+ (1-2\pi + \pi_{00})e^{-8(1-\pi)b/3}. \] (7)

Substituting the limiting distances into (1) gives that, in the limit, \((A, C)\) is preferred to \((A, D)\), since

\[ d_{AD} + d_{CB} - d_{BD} - d_{AC} \to 2a > 0. \]

Substituting the limiting distances into (eq. 1) gives that, in the limit, \((A, C)\) is preferred to \((A, B)\) if

\[ 0 < d_{AB} + d_{CD} - d_{BD} - d_{AC} \]

\[ = 2a - 0.75 \log [E[e^{-4r_{AB}+r_{AC}/3}]/E[e^{-4r_{BC}/3}]]^2. \]

It can be shown that as long as \(\pi_{00} \geq \pi^2\), then

\[ h(b) = 2a - 0.75 \log [E[e^{-4r_{AB}+r_{AC}/3}]/E[e^{-4r_{BC}/3}]]^2 \]

is a decreasing function of \(b\) (for fixed \(a\)). It follows that if a solution, \(h(a)\) of \(h(b) = 0\), can be found, then NJ consistently estimates the topology \((A, C)\) for \(b < h(a)\) and estimates the topology \((A, B)\) for \(b > h(a)\). Examples of zones of consistency for differing choices of \(\pi\) and \(\pi_{00}\) are given in figure 3. With \(\pi_{00} = \pi^2\), the rate variation in the two branches is independent. As \(\pi_{00}\) moves away from \(\pi^2\), making the rates in the two branches more positively correlated, the zone of inconsistency increases.

It is worth noting that the expressions for the limiting distances involve expectations of the form \(E[e^{-4r/b^3}]\). These expectations depend only on the marginal rates-across-sites distribution for \(r\) implied by the parallel rates-across-sites process described here. One might expect that by correctly inferring the marginal rates-across-sites distribution, a correction might be made to avoid inconsistency. Indeed, Tuffley and Steel (1997) prove as much for the covarion rates-across-sites and subtrees model they investigated. For the parallel rates-across-sites process considered here, however, that rate distribution will not be constant across pairs of taxa, as is indicated by equations (6) and (7), and thus correction becomes much more difficult.

### The Archaea-Eukaryotes Split

The split of archaea bacteria from eukaryotes is considered to be near the root of the tree of life, and thus there is considerable interest in which taxa are closest to this split. In the estimated tree of figure 4A the clade of microsporidia (Glugea plecoglossi, Encephalitozoon cuniculi, and Nosema locustae) was placed with a long extending branch at the archaebacteria-eukaryotes split. To illustrate that problems of inconsistency from misspecified distances can occur with a larger number of taxa, we investigate what parameter settings in a model of parallel rate change could cause the microsporidia to be placed at the archaebacteria-eukaryotes split. We do this under the hypothesis that they are actually more closely related to the fungi (Podospora anserina and Schizosaccharomyces pombe in figure 4A). That Microsporidia are truly close relatives of fungi and not deeply branching eukaryotes (as their placement in the elongation factor-1\(\alpha\) would suggest in figure 4A) has gained much recent support from phylogenetic analyses of RNA-encoding and protein genes both in isolation (Keeling and Doolittle 1996; Hirt et al. 1998; Fast, Logsdon, and Doolittle 1999; van de Peer, Ben Ali, and Meyer 2000) and in concatenated sets (Baldauf et al. 2000).

The tree given in figure 4A was constructed with NJ from eukaryotic and archaeabacterial elongation factor-1\(\alpha\) data consisting of 349 sites for 42 taxa generated by modifying Inagaki and Doolittle’s data set (Inagaki and Doolittle 2000). The sequence data for a pair of taxa was used to obtain ML estimates of the distance between them, based on a model using the PAM amino acid substitution matrix, described in Dayhoff and Eck (1968) and Dayhoff, Schwartz, and Orcutt (1979), without a gamma correction; we will refer to them as “uncorrected gamma distances.” The tree that was used in the simulations, with microsporidia more closely related to the fungi, is given in figure 4B. The branch with the microsporidia clade was placed between the fungi and the animals (Homo sapiens and Caenorhabditis elegans in figure 4B), but all other parts of the topology were fixed as in figure 4A. The branch lengths were re-estimated using ML estimation with a fixed topology.

The model for rate change assumes that parallel changes occur along the branch \(A\) leading to the archaebacteria and the branch \(B\) leading to the microsporidia. Specifically, we assume the following:

1. For the branches \(A\) and \(B\), independent of all else, rates are assigned according to table 3.
2. Whenever branches \(A\) and \(B\) are both assigned 0 rates (which happens with probability \(\pi_{00}\)), a 0 rate is assigned to all other branches.
3. Whenever at least one of the rates assigned to \(A\) or \(B\) is non-negative (which happens with probability \(1 - \pi_{00}\)), a common 0 rate is assigned to all other branches with probability \(\pi_{bb}\). Otherwise a common non-zero rate is assigned to all other branches.

The model of rate variation in branches \(A\) and \(B\) is the same model of parallel rate change that gave rise to long branch attraction in the four-taxon case. The main difference between this model and the model in the four-taxon case is that now a large proportion of 0 rate sites are allowed in all other branches of the tree. Under the model above, the expected proportion of sites with 0 rates is \(\pi\) for
A and B, and it is \( p_{00} + p_0^*(1 - p_{00}) \) for all other branches. This provides a simple model for parallel rate variation that could arise because of a loss of function in genes. The cellular structure of microsporidia is highly degenerated, probably because of their parasitic lifestyle. It is also likely that the reduction of these organisms could be extended to the molecular level, since only 1,997 open reading frames appear to be encoded in the \( E.\ cuniculi \) genome (Katinka et al. 2001). Suppose for instance, that some sites in elongation factor 1\( \alpha \) (EF-1\( \alpha \)) are constrained in Eukaryotes but not in archaebacteria because of functions carried out by eukaryotic EF-1\( \alpha \) but never present in archaebacterial EF-1\( \alpha \). If those functions were later lost in microsporidia, the corresponding sites would be free to vary, in contrast to the rest of the eukaryotes and in parallel with the archaebacteria. Eukaryotic EF-1\( \alpha \) (eEF 1\( \alpha \)) proteins are known to have both the core function as a translation elongation factor that is absolutely indispensable for all cellular systems and non-translational (auxiliary) functions that probably are absent in the archaebacterial orthologs (reviewed in Negrutskii and El’skaya 1998). Therefore it is plausible that microsporidian EF-1\( \alpha \) proteins secondarily lost some eukaryotic EF-1\( \alpha \)–specific auxiliary functions during the reductive evolution in this intracellular parasitic lineage (Inagaki et al. 2004). By these parallel absence of EF-1\( \alpha \) auxiliary functions in the microsporidian and archaebacterial proteins, the two sub-trees could show correlated patterns of rate variation, different from other eukaryotes (Inagaki et al. 2004).

To obtain a reasonable set of parameters, \( p_{00}, \pi, \) and \( p_0^* \) for the simulation, we used the proportions of sites with no change among the subgroups of taxa of interest:

<table>
<thead>
<tr>
<th>Group</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaebacteria</td>
<td>0.258</td>
</tr>
<tr>
<td>Eukaryotes</td>
<td>0.436</td>
</tr>
<tr>
<td>Archaebacteria &amp; Microsporidia</td>
<td>0.155</td>
</tr>
</tbody>
</table>

A rough estimate of \( \pi \), the probability of a 0 rate for the archaebacterial subtree, is then 0.258. Similarly 0.155 is an estimate for \( p_{00} \), and 0.436 is an estimate of \( p_{00} + p_0^*(1 - p_{00}) \), the probability of a 0 rate in all other
branches of the tree. Since 0.155 is an estimate for $\pi_{00}$, an estimate for $p_0^*$ can be obtained by solving
\[ 0.155 + p_0^*(1 - 0.155) = 0.436. \]

The solution gives 0.333 as an estimate of $p_0^*$. All of the estimates above will be biased upward because some sites will appear invariant by chance. However, with 42 taxa, the expectation is that the bias will not be large.

The tree used in the simulations is given in figure 4B. The branch with the microsporidia clade was placed between the fungi and the animals, but all other parts of the topology were fixed as in the topology estimated by NJ with uncorrected PAM distances (fig. 4A). If in fact parallel rate variation is present, then uncorrected ML estimates of the branch lengths for the topology in figure 4B will be biased. We used the regression approach described in the Supplementary Material online as a rough bias adjustment.

Because inconsistency implies that wrong topologies will be estimated with high probability when the number of sites is large, we simulated single data sets of 5,000 sites from the model above. Because of the uncertainty in the parameter and branch length estimates, simulations were conducted for a variety of settings for $p_0^*$, $\pi$, and $\pi_{00}$ near the estimated values of 0.333, 0.155, and 0.258. It was clear from these simulations that (1) larger values of the parameters $\pi_{00}$ and $\pi$ are more likely to give rise to microsporidia being placed with the archaea bacteria, (2) changing $p_0^*$ from its estimated value of 0.33 did not change results, and (3) $\pi_{00}$ had to be close to $\pi$ for microsporidia to be placed with the archaea bacteria. The results for some of the parameter settings are indicated in table 4. For the estimated parameters ($p_0^*$, $\pi_{00}$, and $\pi$ set to 0.333, 0.155, and 0.258), the estimated topology places *Entamoeba histolytica* closest to archaea bacteria, as do most of the parameter settings. Parameter settings with $\pi_{00}$ large are the only ones that place microsporidia closest to the archaea bacteria. The estimated topologies did not vary except in the placement of microsporidia. It was of interest to conduct simulations with 349 sites, because this was the size of the data set used. In this case, however, incorrect placement may be explained by sampling variation and,

![Image](https://academic.oup.com/mbe/article-abstract/21/9/1629/992076/1637)

**Table 4**

<table>
<thead>
<tr>
<th>$p_0^*$</th>
<th>$\pi$</th>
<th>$\pi_{00}$</th>
<th>Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>0.26</td>
<td>0.16</td>
<td>E</td>
</tr>
<tr>
<td>0.33</td>
<td>0.26</td>
<td>0.21</td>
<td>E</td>
</tr>
<tr>
<td>0.33</td>
<td>0.26</td>
<td>0.26</td>
<td>M</td>
</tr>
<tr>
<td>0.33</td>
<td>0.31</td>
<td>0.16</td>
<td>E</td>
</tr>
<tr>
<td>0.33</td>
<td>0.31</td>
<td>0.21</td>
<td>E</td>
</tr>
<tr>
<td>0.33</td>
<td>0.31</td>
<td>0.26</td>
<td>M</td>
</tr>
<tr>
<td>0.33</td>
<td>0.31</td>
<td>0.31</td>
<td>M</td>
</tr>
</tbody>
</table>
indeed, the estimated topologies differed more than just in their placement of microsporidia. Nevertheless, the results with respect to the clade closest to the archaebacteria were the same as indicated in table 4.

Although inconsistency did not result at the estimated parameter values, the results in table 4 indicate that there are plausible misspecified distances which could explain the placement of microsporidia near the archaebacteria. The model for parallel rate variation used here is a simple one. Given the results, it is possible that a more refined model would more clearly indicate that the placement of microsporidia can be explained by parallel rate variation. Cases of parallel rate variation may be particularly problematic for “deep” phylogenetic problems, where parallel losses or changes in functional constraints in distantly related lineages are more likely to be observed (Lockhart et al. 1998; 2000).

Regardless of the placement of microsporidia, the example illustrates that results for four taxon trees are likely to have implications for larger number of taxa. In the present case a key feature of the tree used in the simulation was the long branches leading to the microsporidia (0.77) and archaebacteria (0.30). The distance between the nodes corresponding to these branches was not small (0.25), and there were a number of intervening branches but, as indicated in figure 5, the shape of the tree was similar to the shape of the long branch four-taxon trees that have been considered throughout most of this article.

**Transition/Transversion Ratios**

Not adjusting for parameters in the substitution process at a site can lead to long branch inconsistency for NJ. We consider here failure to adjust for a transition/transversion ratio \( R \) different from that implied by the JC model \( (R = 1/2) \). For simplicity we assume a fixed rate model in which case the Kimura 2-parameter model (Kimura 1980) adjusts for a difference. The ML branch length estimates for a pair of taxa under this model is calculated to be

\[
d_{AB} = -\log[1 - 2P - Q]/2 - \log[1 - 2Q],
\]

where \( P \) is the proportion of transition-type differences (the data at a site for the pair of taxa under consideration is \( AG \) or \( CT \)) and \( Q = 1 - P \) is the proportion of transversion-type differences. A plot of the zones of inconsistency when the generating model has a transition/transversion ratio different from 1/2 but a JC model as used as given in figure 6. One can see that the zone of inconsistency increases dramatically as the the transition/transversion ratio increases from 1/2.

**An Example Where Long Branches Repel**

In the examples above and in most of the literature the most common form of inconsistency that arises is long branch attraction. It is, however, possible that long branches will repel. As an example of this phenomenon, we consider a true tree of the form \(((A : b, C : b) : a, (B : a, D : a))\), where the two long branches have length \( b \).
In this case the limiting distances are as follows:

<table>
<thead>
<tr>
<th>Distance</th>
<th>Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d_{ab}(P,0))</td>
<td>(g(2a+b))</td>
</tr>
<tr>
<td>(d_{ad}(P,0))</td>
<td>(g(2a+b))</td>
</tr>
<tr>
<td>(d_{ac}(P,0))</td>
<td>(g(2a+b))</td>
</tr>
<tr>
<td>(d_{bc}(P,0))</td>
<td>(g(2b))</td>
</tr>
<tr>
<td>(d_{bd}(P,0))</td>
<td>(g(2a))</td>
</tr>
</tbody>
</table>

Substituting these limiting distances into equations (1) and (2), we get that, in the limit, \((A, C)\) is preferred to both \((A, B)\) and \((A, D)\), if

\[
2g(b + 2a) - g(2a) - g(2b) > 0. \tag{8}
\]

Suppose now that a JC substitution model is in effect with a gamma rate distribution having \(\alpha = 1\). If the RAS model adopted is incorrectly specified as consisting of a mixture of a fixed rate and a set of invariable sites arising with probability 0.2, then

\[
g(t) = -0.6 \log [1.25(1 + t/0.75)^{-1} - 0.25].
\]

In this case, \(g\) is a convex transformation of \(t\). Whenever this is so, long branch attraction will not occur and there is a possibility that long branches will repel. In this case, taking \(a = 0.25, b = 1.35\) and substituting into (8) we get

\[
2g(b + 2a) - g(2a) - g(2b) = -0.0706 < 0,
\]

and consequently, with probability 1, with large numbers of sites, both \((A, D)\) and \((A, B)\) will be preferred to \((A, C)\).

This is an example of a situation in which NJ does not converge on a particular topology. Because of the symmetry of the problem, with large numbers of sites the probability is approximately 1/2 that the topology \((A, D)\) is estimated and 1/2 of that \((A, B)\) is estimated. In either case, although the generating tree has long branches together, with large numbers of sites, the tree with long branches apart is estimated.

**Connections Between LS, ME, and NJ**

The least-squares estimate of a tree is obtained by choosing a topology and branch lengths so that

\[
\sum_y (d_{ij} - \hat{d}_{ij})^2 \tag{9}
\]

is as small as possible where \(\hat{d}_{ij}\) is the sum of branch lengths along the path between \(i\) and \(j\) in the tree under consideration. This could be unconstrained least-squares estimation—branch lengths are allowed to be negative—or constrained estimation—branch lengths are restricted to be positive. Assuming a true tree \((A : b, C : a) : (A, D), (B : b, D : a))\), and that \(g\) is a monotone concave transformation, a similar analysis to the one for NJ, given in the Supplementary Material online, allows one to draw the following main conclusions.

1. The zone of inconsistency for the unconstrained LS estimate contains the zone of inconsistency for NJ. There are values of \((a, b)\) for which unconstrained LS will be inconsistent and NJ will be consistent. An example is given in figure 7.
Concluding Remarks

The analytical results presented here provide clear, rigorous reasons demonstrating that distance methods become inconsistent and the manner in which they do so when models are misspecified; we expect and simulations suggest that ML methods will exhibit similar zones of inconsistency. It is the shape of the limiting distances, which we have denoted by \( g(t) \) as a function of the true distances \( t \), that determines how and whether there will be a zone of inconsistency. If these functions are concave, there will be a zone of inconsistency and it will favour topologies in which long branches attract. For many of the realistic examples of model misspecification or ignorance discussed here (including the use of Hamming distances), it is clear that, ignoring rate variation, misspecifying rate matrices, \( g(t) \) is a concave function and that long branch attraction results. Indeed, this appeared to be the case for other similar forms of model misspecification considered but not presented here. In light of this tendency it appears reasonable that investigators carefully consider the possibility of model misspecification for topologies with several long branches close to one another. Nonetheless an example was presented in which \( g(t) \) was convex, leading to a long branches repel form of inconsistency. This result indicates that it is not a certainty that long branch attraction will be a consequence of model misspecification. At present we can see no theoretical reasons that other heretofore unconsidered types of model misspecification will most likely lead to concave \( g(t) \) rather than convex \( g(t) \) and consequently long branch repulsion. What is clear through careful consideration of the form of the zone of inconsistency determining equation (4) in the case of long branch attraction and equation (8) in the case of long branch repulsion, is that for a topology to be in the zone, it is necessary for there to be several relatively long branches; this is the only way in which the negative terms in the sum will be larger than the positive terms. The implication is that model misspecification should only be suspected to be a serious potential source of error for estimated topologies with several relatively long branches.

Investigations about inconsistency like the one conducted here provide information about the bias of estimation procedures and not the variance. Through simulation Gaut and Lewis (1995) plot regions where ML estimation performs well with finite samples under model misspecification. For small middle branches \( a \), the shapes of their plots are similar to the shapes given here, suggesting that the poor estimation in these finite sample cases is due in large part to bias. However, for larger middle branches \( a \), the zones where ML performs poorly are much less restricted, indicating, not surprisingly since the variance in distance estimates increases with distance, that variance provides additional problems for estimation in these cases.

For almost all of the examples considered here, it has been shown that inconsistency implies that a particular topology, different from the true one, will be converged on with probability 1. Because a particular topology is being converged on, with a large number of sites, measures of uncertainty, such as for instance bootstrap support, will
suggest that with a high degree of certainty this wrong topology is the only one supported by the data. The fact that a particular topology is being converged on is important for studies investigating the use of simple distance measures as well. An overly simplistic distance measure that has a convex limiting distance function \( g(t) \), for instance, will converge on the wrong topology if the generating topology has long branches together, but it will converge on the right topology if the long branches are apart. In fact, such a distance measure might actually appear to be very efficient in estimating this topology. This is a point that has been considered in some detail in Bruno and Halpern (1999).

The examples involving parallel rate variation were important for several reasons. First, the application to the archaea-bacteria-eukaryotes EF-1\( \alpha \) data indicate that the four-taxon results presented here likely have implications for larger, more complicated topologies. Second, although the model considered is simple, the basic notion of parallel rates-across-sites variation is biologically relevant. Finally, where many of the other forms of model misspecification can be adjusted for with appropriate distances, adjustments here would require a careful consideration of the processes taking place throughout the entire tree. The issues raised are similar to but different from those in Tuffley and Steel (1997). A correction for distances in the covarion model they consider is available as a rates-across-sites adjustment. Similarly, in the parallel rates-across-sites case, corrections could in principle be made through a rates-across-sites adjustment; in contrast, however, the rate distribution will not be constant across the taxa considered. It seems possible that the problem could be diagnosed on the basis of a pilot estimate of the tree, with separate rate distributions being fit to separate subsets of taxa, but we cannot pursue that further here. In any case, some form of iterative distance estimation, iterating between tree estimation and distance estimation with model adjustments, or perhaps ML estimation, would be required.

Acknowledgments

E.S. and A.J.R. are supported by the Natural Sciences and Engineering Research Council of Canada. A.J.R. thanks the Canadian Institute for Advanced Research Program in Evolutionary Biology and the Canadian Institutes for Health Research for fellowship support. A.J.R. and Y.I. were supported by NSERC Operating Grant 227085-00 and NSERC Genomics Grant 228263-99. This collaboration is part of the Prokaryotic Genome Institute for Health Research for fellowship support. E.S. and A.J.R. are supported by the Natural Sciences and Engineering Research Council of Canada. E.S. and A.J.R. are supported by the Natural Sciences and Engineering Research Council of Canada.

Literature Cited


