LETTERS

A Bias in ML Estimates of Branch Lengths in the Presence of Multiple Signals

David Penny, W. T. White, Mike D. Hendy, and Matthew J. Phillips

Allan Wilson Center for Molecular Ecology and Evolution, Massey University, P.O. Box 11222, Palmerston North, New Zealand

Sequence data often have competing signals that are detected by network programs or Lento plots. Such data can be formed by generating sequences on more than one tree, and combining the results, a mixture model. We report that with such mixture models, the estimates of edge (branch) lengths from maximum likelihood (ML) methods that assume a single tree are biased. Based on the observed number of competing signals in real data, such a bias of ML is expected to occur frequently. Because network methods can recover competing signals more accurately, there is a need for ML methods allowing a network. A fundamental problem is that mixture models can have more parameters than can be recovered from the data, so that some mixtures are not, in principle, identifiable. We recommend that network programs be incorporated into best practice analysis, along with ML and Bayesian trees.

Phylogeneticists have been active in finding regions of parameter space where methods for inferring evolutionary trees are, or are not, reliable. Classic simulation studies include Huelsenbeck and Hillis (1993); Gaut and Lewis (1995); and Lockhart et al. (1996). However, most theoretical simulation work has been on a data set generated on a single tree, though we have known for a long time that real sequences have more signals than can fit onto a single tree. This is shown by networks (Bandelt and Dress 1992; Holland et al. 2004; Huson and Bryant 2006), Lento plots (Lento et al. 1995), and the treeness triangle (White et al. 2007).

There has been considerable work on the effects of model violation/misspecification (e.g. Sullivan and Swofford 2001; Buckley 2002), but most work focuses on the Markov model itself and less on whether a single tree is insufficient to describe the data. For the present analysis, we consider a scientific model as consisting of 3 parts (Penny et al. 1992):

- the structure of the model (a tree in our case),
- a mechanism of nucleotide or amino acid change (a stochastic Markov model), and
- the initial conditions (including the nucleotide composition at the root, rates of changes, and the distribution of rates across sites).

Here, we concentrate on cases where model misspecification is the inadequacy of the tree itself, rather than the Markov model.

There are biological examples where mixture models (see Kolaczkowski and Thornton 2004; Matsen and Steel 2007) are biologically reasonable and cases where they are approximating other biological mechanisms. Realistic examples include lineage sorting, hybridization, lateral gene transfer, and with questions of orthology/paralogy among members of a gene family. In other cases, mixture models can mimic other processes, such as similar changes in G + C content leading to convergence onto an incorrect tree (see Lockhart et al. 1992). Thus, we consider that the principles illustrated by mixture models are important in understanding phylogenetic errors.

To illustrate the principles, we generate data on 2 trees, combine the sequences in defined proportions, and test the ability of standard maximum likelihood (ML) to estimate the edge (branch) lengths. Data were generated under a symmetric 2-state Hadamard conjugation (Hendy and Penny 1993) using an Excel spreadsheet (http://awcmee.massey.ac.nz/downloads.htm). This gives exact frequencies of the observed patterns, so there are no sampling errors in either the data or the edge lengths generated from it. We effectively have infinitely long sequences and so concentrate on systematic (Phillips et al. 2004), rather than sampling, errors.

The first simulations used two 4-taxon trees A (1,2)(3,4) and B (1,3)(2,4). In the examples reported here the external branches had 0.2 changes per site and the internal 0.05; only one parameter (the internal edge) is being changed. The 11 data sets had from 100% to 0% of patterns from tree A, decreasing in steps of 10%, with the remaining data coming from tree B.

PAUP* (Swofford 2001) was used on all data sets to estimate the ML values for both trees and branch lengths (4 external and one internal). The ML values for trees are in figure 1A and the branch lengths in figure 1B. As expected, ML favors tree A or B, depending on the proportion of data from each tree, and has equal ML values when the data is mixed 50:50. It is interesting that the ML value for either tree changed little with up to 30% of sites coming from the alternative tree—even though the ML value for the minor tree was increasing. As expected, there was no change in the ML value of the third possible tree (1,4)(2,3), which was not used in generating the data. Because PAUP* requires integer site weights for likelihood computation, the site probabilities were multiplied by 30,001 and rounded to the nearest integer.

Figure 1B has the estimates of branch lengths. The 100% A and 100% B results are important controls, in that branch lengths are estimated correctly (0.2 for the external and 0.05 for the internal). However, with the mixtures, the internal edge is underestimated and the external overestimated. The effect on the external edges is straightforward; the values increase above 0.2, with a maximum overestimate of 6% in the case of the 50:50 mixture.
In this example, the effects are stronger on the internal branch, the estimate being over 30% too low, even considering just one of the 2 competing signals. The expected lengths of the 2 competing signals are both 0.025, and the estimate is 30% lower. However, if the estimate is expressed as a percentage of the sum of the 2 competing signals, we get a 66% underestimate; it is only 34% of the sum. Either way there is a severe underestimate of the length of the internal edge, which will be important when estimating dates of divergence.

There is additional information in the data that allows good recovery of both signals; the problem is forcing ML onto a single tree. The Hadamard (Hendy and Penny 1993), given data from a 50:50 mixture, recovers the 2 signals, each with a value of 0.025 (fig. 1C). The recovery is not exact, in that there is also a very small negative value (-0.000625) for tree C (1,4)(2,3). This exceeds the rounding errors from the computer storage of real numbers, which is $<10^{-16}$ in these calculations. Other network programs can recover signal for both trees as shown in figure 1D for SpectroNet (Huber et al. 2002) for the same 50:50 mixture.

The bias in ML is not inherent to ML but is a form of model violation from forcing the calculation onto a single tree. To show this, we tested a variant of ML with $10^6$ sites from a mixture of the 2 trees. For each of the 11 data sets (fig. 1), each site was assigned to tree A or B with random starting probabilities $P$ and $1 - P$; the trees and edge lengths were given a priori. The search uses a simple hill-climbing algorithm and tests new $P$ values, accepting any which lead to a better log-likelihood score. The value of $P$ that maximizes the log likelihood is selected. With just a single parameter to estimate, this procedure quickly converged to the correct $P$ value, always resulting in a higher log-likelihood value than that produced by fitting either tree alone. Thus, there is additional information in the data that is not used by standard ML on a single tree and which can be used to help detect model violation.

We also did tests on 5-taxon trees. From figure 1B, we can imagine that the additional signal for the internal branch of tree B, which does not fit on tree A, might increase the lengths of edges 1 and 3, or 2 and 4, thus increasing their estimated lengths. To test this, we used 3 unrooted trees on 5 taxa: tree A (1,2)(3,4) and (1,3)(2,4). The highest likelihoods are at the top. B. ML edge estimates for the same data as figure 1A. The external edges are overestimated and the internal edge underestimated. C. The Hadamard conjugation recovers the 2 signals from a 50:50 mixture of data from the 2 trees; there is a very small negative signal for the third tree. The back values are the observed frequencies of the 8 possible patterns in the data; the front values are after correcting for inferred multiple changes at a site. The 2 solid arrows indicate the signals for the 2 trees in the mixture; the dashed arrow is for the potential third tree that is not in the mixture. D. SpectroNet similarly recovers the 2 internal edges from the mixture and represents them as a network.
edges are seriously underestimated, similar to figure 1B. Again, as in figure 1C, we used the Hadamard to recover the signals for both trees from both mixtures. For these examples, the results are accurate to the 5th decimal place. Thus, the failure of ML on a single tree to recover the signals more accurately is not a lack of signal in the data; rather it is model misspecification.

In addition to the model misspecification, there are the well-known problems associated with increased parameterization (see, Steel 2005). With mixture models, there can be more parameters than independent signals in the data. With a standard mixture model on 4 taxa (as in Matsen and Steel 2007), there are a minimum of 5 parameters for the edges of each tree, and one for the proportion of sites from each tree. Thus, there are 11 parameters for the model, which is more than the 7 independent values being estimated from the data—there are 8 patterns in the data for the 2-state symmetric model (Hendy and Penny 1993), but their frequencies must sum to 1.0. The numbers of patterns does increase rapidly with 4-state characters (Hendy et al. 1994), but this is not sufficient to guarantee identifiability. With mixture models, there is the question of identifiability (Chang 1996; Rannala 2002): whether there is in principle sufficient information in the data to uniquely identify the model. Note that in the Matsen and Steel (2007) example, the mixture of 2 sets of edge lengths on a single tree would be rejected by methods that weight against increasing the number of parameters. The mixture requires 6 additional parameters over the alternative single tree that generates the same data.

The problem is exacerbated by the recent tendency to increase the number of parameters being used in models—whether it be from variation in population size (Drummond et al. 2005), recombination (Martin et al. 2005), multiple ML optima (Chor et al. 2000), or different gene trees (Degnan and Rosenberg 2006, see also Stefankovic and Vigoda 2007). These recognize that additional parameters are leading to alternative models giving the same data, and so formal analysis on identifiability of models is urgently required (see Allman and Rhodes 2006 and Matsen and Steel 2007).

Although we expect that with real data, ML on single trees will be biased in estimating branch lengths, this should be kept in perspective. Researchers did not stop searching the space of trees when Graham and Foulds (1982) showed that the problem of evaluating all trees was NP-hard (and therefore cannot be computed exactly as the number of taxa increases). Rather, more care was taken over the search procedures in order to get the best results. We urgently need more work on ML and Bayesian methods for networks (see Strimmer and Moulton 2000; Pagel and Meade 2004). Similarly, we conclude that “best practice” in an analysis is to include a network diagram along with trees from ML and Bayesian analysis.

### Literature Cited


Graham RL, Foulds LR. 1982. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. Math Biosci. 60:133–142.


Martin Embley, Associate Editor.

Accepted November 19, 2007