LETTER

Likelihood, Parsimony, and Heterogeneous Evolution

Matthew Spencer,*† Edward Susko,* and Andrew J. Roger†

*Department of Mathematics and Statistics and †Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Evolutionary rates vary among sites and across the phylogenetic tree (heterotachy). A recent analysis suggested that parsimony can be better than standard likelihood at recovering the true tree given heterotachy. The authors recommended that results from parsimony, which they consider to be nonparametric, be reported alongside likelihood results. They also proposed a mixture model, which was inconsistent but better than either parsimony or standard likelihood under heterotachy. We show that their main conclusion is limited to a special case for the type of model they study. Their mixture model was inconsistent because it was incorrectly implemented. A useful nonparametric model should perform well over a wide range of possible evolutionary models, but parsimony does not have this property. Likelihood-based methods are therefore the best way to deal with heterotachy.

Introduction

Heterotachy is a general term for within-site rate variation over time (Lopez, Casane, and Philippe 2002). Under heterotachy, evolutionary rates at different sites may vary in different ways over subtrees. Such variation is widespread (e.g., Fitch and Markowitz 1970; Uzzell and Corbin 1971; Lopez, Casane, and Philippe 2002; Ané et al. 2005) and can cause biased tree estimation (Lockhart et al. 1998; Inagaki et al. 2004; Susko, Inagaki, and Roger 2004). Models of concomitantly variable codons or nucleotides (covarions or covariotides: Fitch and Markowitz 1970; Galtier 2001; Huelsenbeck 2002), in which sites have constant probabilities over time of switching between two or more rate categories, are a special case of heterotachy. Kolaczkowski and Thornton (2004) (K&T) described an intriguing case where parsimony outperforms misspecified likelihood-based phylogenetic methods under one model of heterotachy. This occurs when the data are divided into partitions with the same tree topology but different edge lengths. Such a situation is not resolved by standard covarion models (e.g., Huelsenbeck 2002), which may represent only a small proportion of possible heterotachous situations (Steel, Huson, and Lockhart 2000; Lopez, Casane, and Philippe 2002). K&T reported that “maximum parsimony performs substantially better than current parametric methods over a wide range of conditions tested.” A Bayesian mixture model with partitions performed better than parsimony and standard likelihood but remained inconsistent. K&T recommended “reporting nonparametric analyses [parsimony] along with parametric results and interpreting likelihood-based inferences with the same caution now applied to two or maximum parsimony trees.” This challenges the widespread belief that parsimony is inferior to likelihood. K&T’s conclusions stem from special choices of conditions and incorrect implementation of likelihood methods. We describe a correct mixture model for partitioned data, and suggest that parsimony does not meet the usual requirements for a nonparametric method.

K&T studied four-taxon trees with two long and two short terminal edges in each partition. In this setting of simple heterotachy models, there are 6 ways to assign two long and two short terminal edges on a labeled four-taxon tree and 15 combinations of two different edge-length partitions. K&T described one such combination (patterns 1 and 5 in fig. 1). Over all combinations (fig. 1), there are nine where both standard likelihood and parsimony perform well. In two cases, both methods perform poorly, but parsimony does slightly better. In four cases, likelihood does better by roughly the same margin. Therefore, likelihood is as good as or better than parsimony in the majority of combinations for the type of mixture model studied by K&T.

K&T simulated evolution using the Jukes-Cantor model (Swofford et al. 1996), which, like parsimony, treats nucleotide substitutions symmetrically. Under the more complex Kimura two-parameter + gamma rates ($K_2P + \Gamma$) model (Swofford et al. 1996), the performance difference between methods is substantially reduced where parsimony outperforms likelihood (e.g., fig. 2a), and remains similar in all other cases (e.g., fig. 2b). Furthermore, evolutionary heterogeneity is unlikely to be as discrete as the two-partition case. A wider distribution of edge lengths is obtained by assigning equal numbers of sites to all six possible edge-length patterns. In this case, for both Jukes-Cantor (fig. 3a) and $K_2P + \Gamma$ (fig. 3b), performance differences are negligible and statistically insignificant.

K&T attribute the poor performance of likelihood-based methods to the nonidentical pattern distribution resulting from assigning edge-length partitions to sites. This attribution is misleading. Edge-length partitions were assigned to sites in a deterministic fashion (edge-length partition $b_1$ to the first half of sites, $b_2$ to the rest), but a randomly selected site is equally likely to have come from either partition. Thus, an appropriate marginal distribution model at a site is a mixture model that assigns probabilities to partitions.

K&T deserve credit for proposing a mixture model Bayesian Markov Chain Monte Carlo with heterotachy (BMCMChetero) that improved on standard likelihood and parsimony methods. K&T weighted likelihood contributions by the posterior probability that the site was in
the partition. In their model, the likelihood contribution for pattern $x_i$ at site $i$ was

$$L(t, b_1, b_2 | x_i) = \rho_{i, t} P(x_i | t, b_1)$$

$$+ (1 - \rho_{i, t}) P(x_i | t, b_2),$$

(1)

where $P(x_i | t, b)$ is the probability of pattern $x$ given tree $t$ and edge-length set $b$. The weight $\rho_{i, t}$ is the posterior probability $P(x_i | t, b_1)/\left[ \sum_{k=1}^{2} P(x_i | t, b_k) \right]$ that $b_1$ is the edge-length partition for $i$. However, this model remained inconsistent. K&T therefore claim that “violating the identical distribution assumption can cause inconsistency, even when the ‘true’ evolutionary model is used.”

This is false. K&T’s model is not a correct likelihood model. The likelihood for the parameters should be the probability of the data given these parameters, so the likelihood contribution for a site is the marginal probability of pattern $x_i$ at $i$

$$L(t, b_1, b_2, \rho | x_i) = \rho P(x_i | t, b_1)$$

$$+ (1 - \rho) P(x_i | t, b_2),$$

(2)

where $\rho$ is the probability that a randomly selected site has edge-length partition $b_1$ (constant across sites). The overall likelihood is obtained by multiplying equation (2) over $i$. This method (fig. 4) performs almost as well as the best-possible case, in which the site partitions and edge-length parameters are known a priori (ML$_{true}$ and BMCMC$_{true}$, Kolaczkowski and Thornton 2004). Mixture models have also recently been used to allow different substitution matrices at different sites (Pagel and Meade 2004) and are likely to become more common in the future.

In the case that sites are independent and have identical distributions, maximum likelihood (ML) estimation will be consistent provided the mixed model satisfies the identifiability condition that incorrect trees do not give the same probabilities of site patterns as the true tree (Chang 1996; Steel and Székely 2002). The claim of K&T that “violating the identical distributions assumption can cause inconsistency” suggests that a problem arises here because edge-length partitions are assigned in a deterministic fashion so that sites in the first partition have a different distribution from sites in the second partition. No matter how edge-length partitions are assigned to sites, as long as fixed proportions of each of them are assigned, ML will be consistent under the same identifiability conditions as in the

\[ \text{FIG. 1.—Performance of parsimony (crosses) and likelihood (circles) over all 15 possible combinations of two different edge-length partitions with two long (0.75 expected changes) and two short (0.05 expected changes) terminal edges, under Jukes-Cantor. The tree diagrams show the six ways to assign two long and two short terminal edges on a four-taxon tree (pattern 1, top left). The panels show the performance of likelihood and parsimony (vertical axis: $f$, proportion of correct tree topologies) as the length of the internal edge (horizontal axis: $r$) increases, for partitions corresponding to the row and column edge-length patterns.} \]
and (for the combination of all six edge-length patterns under \( K2P + \Gamma \)). The main reason for this is that the relative frequencies \( n_x/n \) (where \( n_x \) is the total number of sites) still converge upon the true pattern probabilities \( p(x) \) given in equation (2). Let \( q(x) \) be the probabilities that the ML pattern probabilities, \( q_n(x) \), converge upon as the sequence length gets large. Then the limiting normalized log-likelihoods satisfy

\[
\sum p(x) \log p(x) \geq \sum p(x) \log q(x) = \lim_{\alpha \to 0} \sum p(x) \log q_n(x) \geq \sum p(x) \log p(x),
\]

where the first inequality is Jensen’s inequality and is strict unless \( p(x) = q(x) \) for all \( x \). Because the left- and right-hand sides of equation (3) are the same, it must indeed be the case that \( p(x) = q(x) \): the ML pattern probabilities converge to the true pattern probabilities. This will happen if the tree, edge-length partitions and \( \rho \) converge upon their true values. If the identifiability condition holds, this is the only way it can happen, and ML is consistent. If instead, a set of trees \( T \) gives the same probabilities for all patterns, the only inference about the tree that could ever be drawn from sequence data is that it is in \( T \). Because the limiting likelihoods are maximized by the true pattern probabilities, statistical tests would be able to make this inference.

K&T state that “non-parametric statistical methods are often applied when the assumptions of parametric techniques are violated.” (see also Sanderson and Kim 2000). This is true, but most such methods perform well under almost all parametric assumptions. Simply not requiring a parametric model is not a sufficient criterion for a satisfactory nonparametric method. For example, 0.2 is an estimator of the mean of a distribution, requiring no parametric assumptions. If the true mean is 0.2, the estimator 0.2 will be unbeatable. With small samples, this estimator will also do well for true means close to 0.2. Nevertheless, it will often do very badly. Parsimony performs badly in many cases (e.g., Felsenstein 2004, pp. 107–121). Thus, finding particular situations in which it does less badly than other methods is not a recommendation for its general use.

Likelihood methods allow comparisons of different models. For example, likelihood ratio tests show that for ribosomal RNA and protein-coding genes, covarion models are better descriptions of the data than models in which rates at sites are constant over time (Galtier 2001; Huelsenbeck 2002). Similar analyses may allow us to test more sophisticated models of heterotachy. In contrast, because parsimony does not use explicit models, it cannot answer mechanistic questions of this kind.

In summary, more thorough explorations of edge-length partition combinations and evolutionary models show that under heterotachy, standard likelihood outperforms parsimony overall. The exceptions occur in special cases with oversimplified models, where both methods perform poorly but parsimony is the least bad. Correct likelihood implementation of heterotachy models is the most promising approach.

**Methods**

We used Seq-Gen 1.3 (Rambaut and Grassly 1997) to generate data (200 replicates, sequence length 10,000 in two equal-sized partitions or 12,000 in six equal-sized partitions). Under \( K2P + \Gamma \), the transition-transversion ratio was 2, and we used a continuous gamma distribution with shape parameter 1. We used PAUP* 4 beta 10 for UNIX (Swofford 2003) to reconstruct trees, with all settings as reported in Kolaczkowski and Thornton (2004). For data simulated under Jukes-Cantor, we used the same model when fitting. Where data were simulated under \( K2P + \Gamma \) with continuous gamma-rate variation, we fitted \( K2P \) with a four-category discrete gamma approximation (shape parameter and transition-transversion ratio estimated from the data). For the mixed model, we maximized the likelihood (eq. 2) for each tree \( t \) over edge-length sets \( h_1, b_2 \) and mixing probability \( \rho \). We used the general constrained optimization algorithm VE11 with default settings, available from the Harwell Subroutine Library Archive (http://hsl.rl.ac.uk/archive/hslarchive). Edge lengths were constrained to be nonnegative, and \( \rho \) was constrained to be between 0 and 1.
Acknowledgments

This work was supported by the Genome Atlantic/Genome Canada Prokaryotic Evolution and Diversity Project. We are very grateful to Bryan Kolaczkowski and Joe Thornton for extensive discussions of their work. We are also grateful to David Bryant, Peter Cordes, Chris Field, Yuji Inagaki, Jessica Leigh, Hervé Philippe, and Alastair Simpson for help and comments and to two anonymous referees for constructive criticism.

Literature Cited


Peter Lockhart, Associate Editor

Accepted February 21, 2005