Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 37 (2005) 928–931
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short communication

A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous

Eric A. Gaucher [a,*], Michael M. Miyamoto [b]

[a] *Foundation for Applied Molecular Evolution, Gainesville, FL 32601, USA*
[b] *Department of Zoology, University of Florida, Gainesville, FL 32611-8525, USA*

Received 11 January 2005; revised 21 March 2005
Available online 4 May 2005

1. Introduction

All methods of phylogenetic inference make assumptions about the underlying evolutionary process of their characters and it is these assumptions that determine their relative successes and failures in the estimation of the true phylogeny for a group (Hillis, 1995). This dependency of phylogenetic accuracy and robustness on evolutionary assumptions has been most extensively studied for the classic case of Felsenstein (1978) and its four-taxon phylogeny with two long, unrelated, terminal branches interspersed with two short ones (Fig. 1A). Given this model phylogeny, "long branch attraction" can occur and thereby lead to the convergence of a phylogenetic method onto an incorrect tree with the two long and two short terminal branches directly connected rather than interspersed. The extent to which a particular phylogenetic method is susceptible to this problem depends on what assumptions it makes about the evolution of the characters and data themselves.

Recently, Kolaczkowski and Thornton (2004) extended this classic problem of long-branch attraction to the more complex situation of partitioned sequences (i.e., those with two separate subsets of sites that are evolving under two opposing sets of long and short terminal branches; Fig. 1B). Their simulations emphasized the most extreme bi-partitioning of the sites (an even 50:50 split). Under these conditions, maximum parsimony (MP) generally outperformed its likelihood-based counterparts (maximum likelihood (ML) and Bayesian phylogenetics) in terms of its phylogenetic accuracy. Thus, in contrast to the classic Felsenstein case, Kolaczkowski and Thornton identified a situation where MP can be generally preferred over ML and Bayesian phylogenetic approaches.

Simulation studies are most powerful and informative when they compare the relative performances of different methods under a broad continuum of conditions (Huelsenbeck, 1995; Huelsenbeck and Hillis, 1993). In light of this fact, our study now presents additional simulations that focus on the relative performances of ML versus MP as one shifts from the simpler Felsenstein heterogeneity to the more complex Kolaczkowski/Thornton heterogeneity. Specifically, our simulations evaluate the robustness or phylogenetic accuracy of different approaches when one or more of their underlying assumptions is violated. With these additional simulations, ML is shown to generally outperform MP even for partitioned sequences. In concert with other theoretical and empirical considerations, these results support a preference for the further development, implementation, and application of likelihood-based approaches even when evolution is heterogeneous.

2. Methods

In contrast to Kolaczkowski and Thornton (2004), the classic Felsenstein (1978) case is referred to herein as "heterogeneous," rather than "homogeneous," in recognition of the fact that at least two major rate shifts have occurred across its model phylogeny (Fig. 1A). Thus, our current use of heterogeneous follows the definition of Lopez et al. (2002, p. 1) for "heterotachy" as "within-site rate variations … throughout time." Furthermore,

* Corresponding author. Fax: +1 352 271 7076.
E-mail address: egaucher@ffame.org (E.A. Gaucher).

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2005.03.027

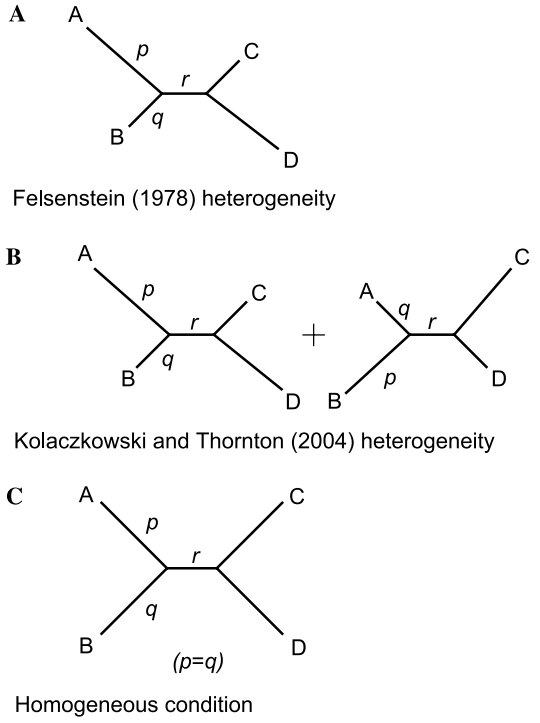

Fig. 1. (A) Felsenstein heterogeneity as illustrated by its model phylogeny with two long versus two short, unrelated, terminal branches. (B) Kolaczkowski/Thornton heterogeneity as illustrated by its two model trees with opposing long and short terminal branches. (C) The homogeneous situation as represented by its model tree with equal terminal branches. p, q, and r refer to the branch lengths for the long terminal, short terminal, and internal branches, respectively.

although Kolaczkowski and Thornton evaluated both ML and Bayesian phylogenetic methods, only ML is considered here in light of their nearly identical results for both of these likelihood-based approaches.

Otherwise, our evolutionary simulations, phylogenetic analyses, and terms followed the general procedures and terminology of these authors. DNA sequences of 10,000 bases each were simulated under the Jukes and Cantor (1969) model with the Evolver program in PAML, version 3.14 (Yang, 1997). Sites were simulated on the two opposing trees in Fig. 1B in proportions of 50:50, 60:40, 70:30, 80:20, 90:10, and 100:0. In these simulations, the lengths of the two long terminal branches were fixed at $p = 0.75$ substitutions/site, whereas those for the short terminal branches were set at $q = 0.05$, 0.15, 0.25, or 0.375 substitutions/site. The length of the internal branch (r) was first set at 0.000, 0.0125, and 0.025, and then varied to 0.400 in increments of 0.025. For the critical 60:40, 70:30, and 80:20 splits, additional simulations were conducted for other values of r to increase the density of data points around their BL_{50} inflections (see below). All conditions were replicated 200 times.

ML and MP analyses of the simulated data sets were conducted with PAUP*, version 4.0b10 (Swofford, 1998). Optimal trees were obtained by exhaustive searching, with the ML analyses relying on the Jukes/Cantor model. The performances of both methods were summarized by their phylogenetic accuracy (i.e., proportion of times out of 200 replicates that the true tree was uniquely recovered). From plots of this metric versus r (Fig. 2A), estimates of the internal branch lengths at 50% accuracy (BL_{50}) were obtained and then statistically compared between ML and MP for the critical 60:40, 70:30, and 80:20 simulations with the PROBIT procedure of SAS/STAT, version 8.00 (SAS Institute Inc., 1999). The additional simulations for these critical splits focused on other values of r taken from around their BL_{50} estimates (e.g., $r = 0.20625$, 0.21250, 0.21875, 0.22500, 0.23125, and 0.23750 for the 70:30 split, $q = 0.05$, and ML).

3. Results and discussion

As previously documented by Kolaczkowski and Thornton (2004), MP approached 100% phylogenetic accuracy faster than ML in all four simulations with a 50:50 partition of sites (Fig. 2A). As ML averages the different p and q branch lengths for the two partitions, it underestimates (not overestimates contrary to their assumed typo on page 982) the internal branch lengths of the true trees and thereby their full support. MP is less sensitive to this heterogeneity since branch lengths are not considered in its phylogenetic evaluations. However, this greater performance was transitory as ML and MP performed similarly in all four simulations with a 70:30 split, with ML then outperforming MP in all of the 80:20, 90:10, and 100:0 trials. The greater performance by ML for the 100:0 simulations is expected as this split corresponds to the classic case of Felsenstein (1978) heterogeneity where the well-known problem of long-branch attraction constitutes less of a concern for ML than for MP. For this same reason, ML also outperformed MP in the 80:20 and 90:10 simulations.

Increasing q from 0.05 to 0.375 resulted in improved performances by both ML and MP (Fig. 2A). These improvements are due to reductions in both Felsenstein and Kolaczkowski/Thornton heterogeneity as q approaches p, thereby converging onto a homogeneous condition (Fig. 1C).

ML outperforms MP over more of the parameter space examined in this study (Fig. 2B). This overall greater performance by ML raises the question as to what extent real sequences are partitioned. Although few such studies exist to answer this question, Kolaczkowski and Thornton cited an intriguing case in which the degree of sequence partitioning appears to be greater than 80:20 (Inagaki et al., 2004). This estimate of >80:20 is derived from a cross comparison of those sites for elongation factor 1α that supported an incorrect tree of Microsporidia with Archaea versus those positions that were evolving significantly faster in either "short-branch" eukaryotes or

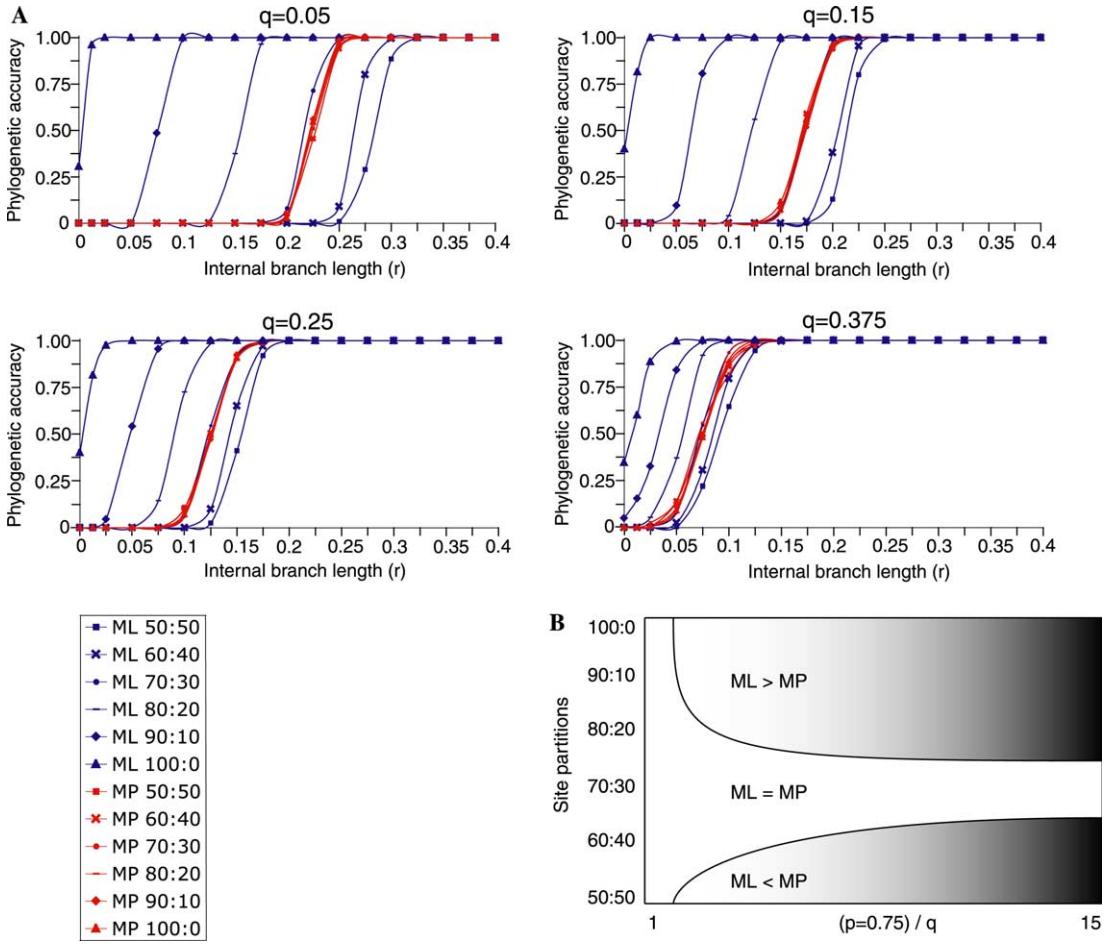

Fig. 2. (A) ML and MP results for the simulations with a

Archaea, but not both (their Figs. 3 and 4). ML analysis by Inagaki et al. (2004) of their "Micro*" data set recovered this incorrect tree, but at a low bootstrap score of only 34% (their Fig. 2A). In contrast, our MP analysis of their Micro* sequences resulted in the same incorrect tree, but with a much greater bootstrap score of 95% (results not shown). Thus, although both ML and MP recover the same incorrect tree as optimal, the former is not as strongly biased as the latter according to their bootstrap scores. This improvement in the relative performance of ML versus MP is expected given that a >80:20 split falls in a region of our parameter space that favors ML over MP (Fig. 2B).

However, perhaps an even stronger reason to continue emphasizing likelihood-based approaches is that they offer the rigorous statistical framework to develop, test, and apply phylogenetic methods intended to capture biologically relevant modes of sequence evolution (Felsenstein, 2004). Included here would be the covarion process (Fitch and Markowitz, 1970) and other recent models of heterotachy as introduced by Tuffley and Steel (1998), Galtier (2001), Penny et al. (2001), Susko et al. (2002), Huelsenbeck (2002), *inter alia* (for review see Gaucher et al., 2002). Along these lines, Kolaczkowski and Thornton presented a new mixture model to account for their partitioned sequences, one that worked extremely well in their simulations. Despite its success, these authors nevertheless were conservative in their promotion of mixture models because of concerns about estimating the most appropriate number of partitions for real sequences and the computational burdens of implementing more complex evolutionary models. However, these concerns can be accommodated in a likelihood-based analysis by, e.g., Bayes factors or likelihood ratio tests for partition estimation and by, e.g., Markov chain Monte Carlo sampling for tractability (Lartillot and Philippe, 2004; Pagel and Meade, 2004). For these multiple reasons, we conclude by calling for once again mixture models and likelihood-based approaches even when evolution is heterogeneous.

Acknowledgments

We thank Tang Li, Slim Sassi, Richard A. Kiltie, and Steve Benner for their assistance with this study; NASA Exobiology for its grant to E.A.G.; and the Department of Zoology, University of Florida for its support to M.M.M.

References

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

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland.

Fitch, W.M., Markowitz, E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4, 579–593.

Galtier, N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18, 866–873.

Gaucher, E.A., Gu, X., Miyamoto, M.M., Benner, S.A., 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. 27, 315–321.

Hillis, D.M., 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44, 3–16.

Huelsenbeck, J.P., 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19, 698–707.

Huelsenbeck, J.P., 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44, 17–48.

Huelsenbeck, J.P., Hillis, D.M., 1993. Success of phylogenetic methods in the 4-taxon case. Syst. Biol. 42, 247–264.

Inagaki, Y., Susko, E., Fast, N.M., Roger, A.J., 2004. Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1alpha phylogenies. Mol. Biol. Evol. 21, 1340–1349.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431, 980–984.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.

Lopez, P., Casane, D., Philippe, H., 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19, 1–7.

Pagel, M., Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53, 571–581.

Penny, D., McComish, B.J., Charleston, M.A., Hendy, M.D., 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. 53, 711–723.

SAS Institute Inc., 1999. SAS/STAT, version 8.00. Cary, NC.

Susko, E., Inagaki, Y., Field, C., Holder, M.E., Roger, A.J., 2002. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol. Biol. Evol. 19, 1514–1523.

Swofford, D.L., 1998. PAUP 4.0*— Phylogenetic Analysis Using Parsimony (* and Other Methods). Sinauer Associates, Sunderland.

Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147, 63–91.

Yang, Z.H., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555–556.