COMMENTS

Evolution, 55(6), 2001, pp. 1261–1263

POLYTOMIES, THE POWER OF PHYLOGENETIC INFERENCE, AND THE STOCHASTIC NATURE OF MOLECULAR EVOLUTION: A COMMENT ON WALSH ET AL. (1999)

Edward L. Braun[1] and Rebecca T. Kimball[2]
[1]*Department of Plant Biology, The Ohio State University, Columbus, Ohio 43210*
E-mail: braun.83@osu.edu
[2]*Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, Ohio 43210*

Key words.—Molecular evolution, phylogeny, polytomy, sample size, speciation, statistical power analysis.

Received March 28, 2000. Accepted November 9, 2000.

Walsh et al. (1999) recently suggested an innovative application of power analysis (Cohen 1977) to explore the nature of polytomies (multifurcating rather than bifurcating relationships) in phylogenetic inference. The frequent recovery of polytomies in phylogenetic analyses has prompted substantial interest in the underlying biological nature of these inferred polytomies. Many authors assume that polytomies simply reflect the inability to resolve bifurcating relationships (''soft'' polytomies) and suggest that additional data or improved analyses will allow the recovery of the underlying relationships (e.g., Maddison 1989; DeSalle et al. 1994). In contrast, other researchers consider polytomous relationships to be valid phylogenetic hypotheses reflecting multiple simultaneous speciation events (a ''hard'' polytomy; Hoelzer and Melnick 1994a,b). Examining the differences between these alternative viewpoints is extremely difficult, because soft polytomies do not exhibit clear differences from hard polytomies in phylogenetic analyses.

Specifically, Walsh et al. (1999) sought to determine whether internal branch lengths which were not significantly different from zero simply reflect an insufficient sample size (amount of sequence data). The example developed by Walsh et al. (1999) involved the polytomous relationships among auklets (Charadriiformes: Alcidae) inferred using mitochondrial DNA sequences. Since this inferred polytomy might reflect a soft polytomy (defined by Walsh et al. [1999] as speciation during successive glacial/interglacial periods during the late Pliocene and early Pleistocene) or a hard polytomy (defined as multiple speciations during the same period of climatic oscillation), Walsh et al. (1999) used power analysis to determine whether the number of base pairs (sites) of sequence data obtained was sufficiently large to detect substitutions along the internal branches if the speciation was not simultaneous (a soft polytomy). We found this an innovative application of classical power analysis and an excellent alternative to Monte Carlo simulation (e.g., Saitou and Nei 1986; Hillis et al. 1994; Huelsenbeck et al. 1996).

We believe that an extremely important aspect of the work by Walsh et al. (1999) is their clear statement of the alternative and null hypotheses concerning the differentiation between hard and soft polytomies. As they point out, ''resolution of the biological reality of polytomies is complicated by the fact that a polytomy represents the null hypothesis for phylogenetic reconstruction—all taxa are equally related—and therefore cannot be proven.'' By restating the problem in the context of the power necessary to detect short internal branches, Walsh et al. (1999) place the hypothesis of a hard polytomy in a much more testable context. However, we were surprised to find that power analysis suggests only 215 to 1237 base pairs of mitochondrial sequence data are required for resolution of a soft polytomy in the auklets, as defined by Walsh et al. (1999).

To further explore the study of polytomies in a statistical context, we examined the implications of the inherent variance exhibited by the nucleotide substitution process. It is clear that molecular evolution is a stochastic process with the number, type, and position of fixed sequence differences being determined by the processes of mutation, selection, and genetic drift. The stochastic nature of molecular evolution may have a profound impact upon the number of differences observed during a specific time period and it is possible that many fewer (or many more) mutations will be fixed during a particular period of time than one expects based upon the total number of substitutions in the phylogenetic tree.

Nucleotide substitution can be modeled as a Poisson process (e.g., Wilson et al. 1987), with the probability (P) of N nucleotide substitutions occurring during a given period of time given by:

$$P = (e^{-\mu}\mu^{N})/N! \qquad (1)$$

where μ, the mean expected number of nucleotide substitutions during a time period, is defined as:

$$\mu = L\lambda t \qquad (2)$$

In equation (2), L is the sequence length (in base pairs), λ is the mean rate of nucleotide substitution (in substitutions per site per year), and t is the time period considered (in years). If one uses the estimate of evolutionary rate presented by Walsh et al. (1999) (1.35×10^{-8} substitutions per site per year) and the largest sample size estimated by Walsh et al. (1237 base pairs), the mean number of expected substitutions (μ) during 100,000 years is 1.67. However, based upon equation (1), there is an 18.8 % probability of observing no substitutions in that time period. In fact, using 215 base pairs, the smallest sample size proposed by Walsh et al. (1999), $\mu = 0.29$ and there is a 74.8% probability of observing no substitutions at all.

The use of a Poisson model of substitution suggests an alternative approach to assessing the sample size necessary to see an effect, which we will define here as the presence of at least one substitution along an internal branch. From equations (1) and (2) one can determine the sequence length (L) necessary to be 95% confident that at least one substitution will have occurred along a branch:

$$L = -\log \text{normal}(0.05)/\lambda t \qquad (3)$$

(other confidence intervals could be specified simply by changing the numerator of equation 3). For the auklet problem, at least 2219 bp of mitochondrial sequence data are required before one can be 95% confident that at least one substitution will have occurred during a period of 100,000 years. Thus, our results imply that the stochastic nature of molecular evolution reduces the power of phylogenetic analysis quite substantially, and one must obtain sequence information for a greater number of mitochondrial nucleotide sites than implied by the results presented in Walsh et al. (1999).

Considering the length of typical avian mitochondrial genomes (~16 kb) and assuming that λ is constant across the mitochondrial genome, the shortest period of time for which one can be 95% confident of observing a substitution is approximately 14,000 years, even if complete mitochondrial DNA sequences are obtained. Assuming that λ is constant across the mitochondrial genome may seem unrealistic given a number of analyses documenting substantial differences in the mean rate of evolution for different mitochondrial genes as well as substantial among-site rate variation (e.g., Kumar 1996). However, studies that have documented substantial rate variation in mitochondrial genomes have typically considered both synonymous and nonsynonymous sites, while studies which have considered silent substitution rates have concluded that there is little variation in the rate of evolution for these sites across vertebrate mitochondrial genomes (Nedbal and Flynn 1998, and references cited therein). Since mutations at synonymous sites predominate in comparisons of recently diverged taxa, the impact of rate variation across the mitochondrial genome is likely to be fairly limited for problems similar to the auklet example discussed by Walsh et al. (1999).

Because synonymous sites are less affected by rate variation, we explored a simple extension of the method proposed here by considering only synonymous substitutions. Examination of the complete *Gallus gallus* mitochondrial DNA sequence (Desjardins and Morais 1990) indicates that 39 % of sites in protein coding regions are synonymous positions. One can then estimate the number of synonymous sites necessary to be 95 % confident that at least one substitution has occurred and then use the proportion of synonymous sites to extrapolate the total number sites necessary. Using cytochrome *b* sequence data from a recent avian phylogenetic study (Kimball et al. 1999), we estimated that the mean rate of synonymous substitutions is $\sim 3.5 \times 10^{-8}$ substitutions per synonymous site per year, similar to estimates obtained in other vertebrate lineages (Brown et al. 1982). Thus, if one assumes that there is little among-site rate variation in synonymous sites (making the Poisson model appropriate) and that substitutions at most nonsynonymous sites are constrained, 856 synonymous sites are sufficient to be 95 % confident that at least one synonymous substitution would have occurred during a 100,000 year period. This corresponds to 2195 base pairs of mitochondrial coding sequence, which is very similar to the estimate presented above that assumed a simpler model of evolution.

Although this more complex model of sequence evolution had a modest impact upon the apparent power of phylogenetic analyses to resolve soft polytomies, the more complex model does have profound implications regarding the distribution of homoplasy. This can be illustrated by considering the probability that a single nucleotide change uniting two taxa is obscured by a subsequent change in one taxon, reflecting either the reversal of that change or a change to another nucleotide obscuring the synapomorphy. For the auklet problem, the terminal branches correspond to ~2.6 million years (Walsh et al. 1999). Using the values calculated by Walsh et al. (1999) and assuming that λ is constant across the sequences, the probability that a single synapomorphy will be unaltered is 93.2%. In contrast, the probability that a single synapomorphy at a more rapidly evolving synonymous site will be unaffected by subsequent substitutions is 83.4%. Thus, more realistic models of evolution that concentrate changes in a smaller set of rapidly evolving sites, such as the synonymous sites, will be more affected by homoplasy. Since the impact of homoplasy upon inferences regarding the nature of a polytomy would be affected by the number of taxa examined and the distance to any outgroup sequences included, there is no simple method to incorporate its effects into the method we propose here. However, the use of Monte Carlo simulation may ultimately prove to be an appropriate method to examine polytomies in these more complex situations.

Regardless of the potential impact of homoplasy, we believe that our equation (3) can provide useful information. When homoplasy has had little impact upon the data, it can provide a realistic estimate of the sample size necessary to differentiate between soft and hard polytomies. Since homoplasy acts to degrade phylogenetic information, our proposed method also provides a useful minimum sample size for cases where homoplasy has had a significant impact upon the data. One problem with this approach is its potential to motivate the collection of large amounts of additional data, which could potentially converge upon an incorrect estimate of relationships (such as data that exhibit the long-branch attraction phenomenon described by Felsenstein 1978). However, this concern is not relevant when analysis are undertaken *a posteriori,* such as the case described by Walsh et al. (1999). In such a situation, this method just allows an assessment of the confidence which should be placed upon short branches (i.e., if the probability of even a single nucleotide substitution along a branch is below a specified value, such branches should be viewed with skepticism).

It should be possible to apply this approach in general to sequence data, if biologically reasonable values for the evolutionary rate and internode divergence time could be obtained. A commonly used estimate for the mean rate of substitution (λ) in vertebrate mitochondrial DNA is 10^{-8} substitutions per site per year (which corresponds to 2% divergence per million years; Brown 1983). This value has been independently estimated in a variety of avian taxa (Klicka and Zink 1997) and may be reasonable to use for avian mitochondrial DNA when an independent estimate of λ is not available. Likewise, the rate of 3.5×10^{-8} substitutions per synonymous site per year is likely to represent a reasonable estimate for avian taxa if one focuses upon silent changes. Estimates for internode divergence time can be obtained from biogeographic or climatic changes which may be associated with the relevant speciation events. Alternatively, other information such as typical intraspecific coalescence times (e.g., Moore 1995) could be used.

Clearly, it is possible for analyses of this type to be confounded by the variation of rates among different lineages. There is ample evidence for the variation of both synonymous and nonsynonymous rates among taxa, although most studies have shown that there is less variation in synonymous rates (e.g., Muse and Gaut 1997). Although methods to better understand the impact of rate variation among lineages may be extremely useful for these types of analyses, the majority of evidence suggests that the rates similar to those used in this study are appropriate for the majority of avian taxa (see Klicka and Zink 1997 and references cited therein). Indeed, these rate estimates may be especially helpful for use *a priori* during experimental design, with refined values estimated directly from the data being used whenever possible for analyses conducted *a posteriori*.

The potential for bias in the estimation of internal branch lengths due to homoplasy—especially if one considers more complex models of evolution characterized by among-site rate variation—suggests that the method we propose will primarily be useful for recent divergences, such as the auklet radiation which is thought to have occurred ~2.6 million years ago (Walsh et al. 1999). Since we believe that the minimum amount of data required to resolve a soft polytomy must imply a reasonable likelihood that synapomorphies uniting relevant taxa exist, any method for determining the power of phylogenetic analyses under these conditions should find sample sizes at least as large as those implied by our equation (3). Like other problems in phylogenetics, such analyses may ultimately reveal that specific questions cannot be adequately resolved. Despite these problems, we are confident that further examination of the underlying biology of hard polytomies will provide insights into evolutionary processes, and the approaches suggested in this paper and by Walsh et al. (1999) should provide an excellent foundation for identification of hard polytomies in future analyses.

Acknowledgments

We thank H. E. Walsh and colleagues for stimulating this work. Two anonymous reviewers provided comments that greatly improved this manuscript. E.L.B. was supported by a United States Department of Agriculture fellowship (1999-01582) and R.T.K. was supported by a National Science Foundation grant to D. J. Crawford.

Literature Cited

Brown, W. M. 1983. Evolution of animal mitochondrial DNA. Pp. 62–88 *in* M. Nei and R. K. Koehn, eds. Evolution of genes and proteins. Sinauer, Sunderland, MA.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18:225–239.

Cohen, J. 1977. Statistical power analysis for the behavioral sciences. Academic Press, New York.

DeSalle, R., R. Absher, and G. Amato. 1994. Speciation and phylogenetic resolution. Trends Ecol. Evol. 9:297–298.

Desjardins, P., and R. Morais. 1990. Sequence and gene organization of the chicken mitochondrial genome: a novel gene order in higher vertebrates. J. Mol. Biol. 212:599–634.

Felsenstein, J. 1978. Cases in which parsimony and compatability methods will be positively misleading. Syst. Zool. 27:27–33.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science 264:671–677.

Hoelzer, G. A., and D. J. Melnick. 1994a. Patterns of speciation and limits to phylogenetic resolution. Trends Ecol. Evol. 9: 104–107.

———. 1994b. Reply from G. A. Hoelzer and D. J. Melnick. Trends Ecol. Evol. 9:298–299.

Huelsenbeck, J. P., D. M. Hillis, and R. Jones. 1996. Parametric bootstrapping in molecular phylogenies: applications and performance. Pp. 19–45 *in* J. D. Ferraris and S. R. Palumbi, eds. Molecular zoology: advances, strategies, and protocols. Wiley, New York.

Kimball, R. T., E. L. Braun, P. W. Zwartjes, T. M. Crowe, and J. D. Ligon. 1999. A molecular phylogeny of the pheasants and partridges suggests these lineages are polyphyletic. Mol. Phylogenet. Evol. 11:38–54.

Klicka, J., and R. M. Zink. 1997. The importance of recent ice ages in speciation: a failed paradigm. Science 277:1666–1669.

Kumar, S. 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. Genetics 143:537–548.

Maddison, W. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377.

Moore, W. S. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution 49:718–726.

Muse, S. V., and B. S. Gaut. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. Genetics 146:393–399.

Nedbal, M. A., and J. J. Flynn. 1998. Do the combined effects of the asymmetric process of replication and DNA damage from oxygen radicals produce a mutation-rate signature in the mitochondrial genome? Mol. Biol. Evol. 15:219–223.

Saitou, N., and M. Nei. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. 24:189–204.

Walsh, H. E., M. G. Kidd, T. Moum, and V. L. Friesen. 1999. Polytomies and the power of phylogenetic inference. Evolution 53:932–937.

Wilson, A. C., H. Ochman, and E. M. Prager. 1987. Molecular time scale for evolution. Trends Genet. 3:241–247.

Corresponding Editor: M. Zelditch