Is the Felsenstein Zone a Fly Trap?

John P. Huelsenbeck

Systematic Biology, Vol. 46, No. 1 (Mar., 1997), 69-74.

Stable URL:
http://links.jstor.org/sici?sici=1063-5157%28199703%2946%3A1%3C69%3AITFZAF%3E2.0.CO%3B2-W

Systematic Biology is currently published by Society of Systematic Biologists.

Syst. Biol. 46(1):69–74, 1997

IS THE FELSENSTEIN ZONE A FLY TRAP?

JOHN P. HUELSENBECK

Department of Integrative Biology, University of California, Berkeley, California 94720, USA; E-mail: johnh@mws4.biol.berkeley.edu

Abstract.—Although long-branch attraction, the incorrect grouping of long lineages in a phylogeny because of systematic error, has been identified as a potential source of error in phylogenetic analysis for almost two decades, no empirical examples of the phenomenon exist. Here, I outline several criteria for identifying long-branch attraction and apply these criteria to 18S ribosomal DNA (rDNA) sequence data for 13 insects. Parsimony and minimum evolution with p distances group the two longest branches together (those leading to Strepsiptera and Diptera). Simulation studies show that the long branches are long enough to attract. When a tree is assumed in which Strepsiptera and Diptera are separated and many data sets are simulated for that tree (using the parameter estimates for that tree for the original data), parsimony analysis of the simulated data consistently groups Strepsiptera and Diptera. Analyses of the 18S rDNA sequences using methods that are less sensitive to the problem of long-branch attraction estimate trees in which the long branches are separate. [Felsenstein zone; insect phylogeny; long-branch attraction; neighbor joining; parsimony.]

Phylogenetic methods can become inconsistent, i.e., converge to an incorrect genealogical tree as more data are added, when the assumptions of the method are severely violated (Felsenstein, 1978; Hendy, and Penny, 1989). The combination of evolutionary parameters for which a method will provide inconsistent estimates of phylogeny has been termed the Felsenstein zone (Huelsenbeck and Hillis, 1993). Felsenstein first pointed out that the parsimony method will converge to a phylogenetic estimate in which long branches are linked together when, in reality, the long branches are separated by very short branches, hence the maxim that "long branches attract" (Felsenstein, 1978). Although inconsistency has been identified through theoretical studies as a potential problem, no convincing examples exist that suggest that inconsistency may be a problem for phylogenetic methods in nature.

Carmean and Crespi (1995) suggested that the maximum parsimony method has converged to an incorrect estimate of a Strepsiptera + Diptera (flies) relationship because of the inconsistency problem. Strepsiptera are a moderately speciose group (532 described species; Kathirithamby, 1991) of parasitic insects that have been traditionally placed within Coleoptera (Crowson, 1960, 1981; Arnett, 1968; Ross et al., 1982) or as the sister taxon of Coleoptera (Kathirithamby, 1991; Kristensen, 1991). Carmean and Crespi (1995) based their conclusion on the facts that (1) the branches leading to Strepsiptera and Diptera are both very long and (2) the support for this grouping is moderately high according to the bootstrap method. Unfortunately, these criteria for long-branch attraction are weak because they fail to identify whether the branches are long enough to attract each other in a parsimony analysis. According to these criteria, if the longest branches of a tree happen to be linked together, then long-branch attraction or method inconsistency can be invoked. Yet, using these criteria, it is impossible to ascertain whether (1) long branches do, in fact, belong together or whether (2) the long branches should be separated by short branches but were linked together because of long-branch attraction.

I argue that two more tests must be passed before long-branch attraction can be invoked: it should also be shown (1) that the branches are long enough to attract (i.e., if the long branches were separated, that the maximum parsimony method would link them together in the estimated phylogeny) and (2) that a meth-

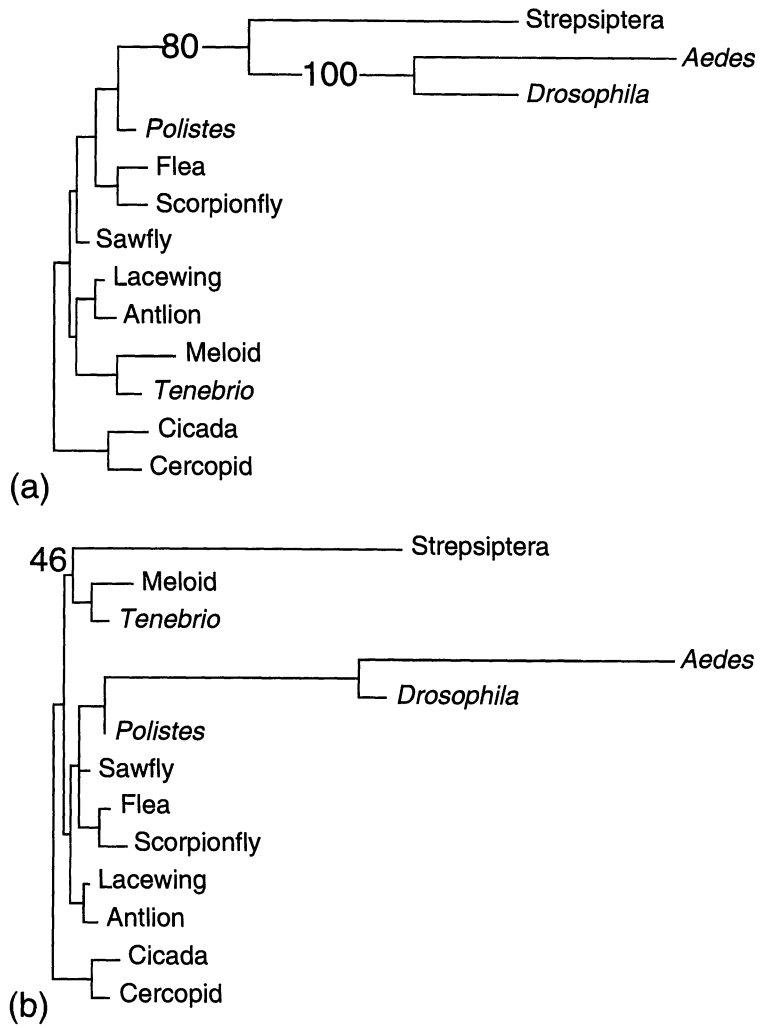

FIGURE 1. Phylogenetic trees for 13 insect groups. Numbers on branches are bootstrap values. (a) One of the trees estimated using the maximum parsimony method with Fitch (1971) optimization. Maximum parsimony estimated 27 trees; each was 364 steps in length. (b) Tree estimated using the maximum likelihood method implemented with the HKY85+Γ_5 model of DNA substitution (Hasegawa et al., 1985; Yang, 1993). The log likelihood of this tree is −2822.86, and the maximum likelihood estimates of κ and α are 3.60 and 0.29, respectively.

od that is less sensitive to the long-branch attraction problem gives a phylogenetic estimate in which the long branches are separated.

ANALYSIS

I performed additional analyses of the 18S ribosomal DNA (rDNA) data for 13 holometabolous insects. The alignment of 18S rDNA sequences from Carmean and Crespi (1995) was used with the exception that all sites with gaps, missing data, or an ambiguous alignment were removed (GenBank accession numbers: M21017, U06478, U06480, X07801, X57172, X77784, X77785, X77786, Z26765). The alignment is available from http://mw511.biol.berkeley.edu/john/john.html.

Figure 1a shows a tree estimated by the maximum parsimony method assuming

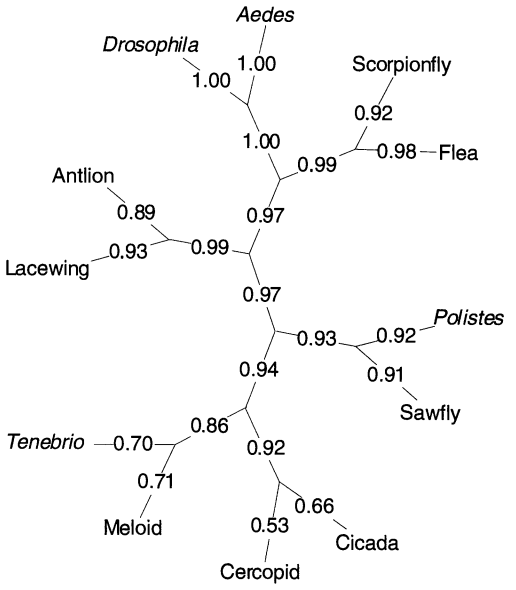

FIGURE 2. Strepsiptera and Diptera grouped together in maximum parsimony analyses even if the lineages were separated in the simulated tree. The number on each branch represents the proportion of the time that a Stepsiptera + Diptera grouping was estimated using parsimony even if Strepsiptera were joined to the branch. The proportions are based on simulated data sets of the same size as the original 18S rDNA data. However, the asymptotic behavior of the maximum parsimony method is to join Stepsiptera and Diptera regardless of where Strepsiptera are placed on the tree.

Fitch (1971) optimization. Maximum parsimony estimated 27 trees, each 364 steps in length, of which the tree of Figure 1a is one. As noted by Carmean and Crespi, the branches leading to Strepsiptera and Diptera (represented by *Aedes* and *Drosophila*) are the longest, and the bootstrap support for this grouping is moderately high (80% bootstrap support for the Strepsiptera + *Aedes* + *Drosophila* clade; Efron, 1979, 1982; Felsenstein, 1985). The bootstrap support for all but the lacewing + antlion clade on the tree were low (<75%). This tree is similar to the parsimony tree estimated by Carmean and Crespi (1995). The other 26 equally parsimonious trees (not shown) are similar in that Strepsiptera and Diptera are monophyletic. The best tree for which Strepsiptera and Diptera are not monophyletic is three steps shorter (length = 367 steps) but is significantly different according to an a priori T-PTP test ($P < 0.01$; Faith, 1991).

I performed simulation analyses to test whether the branches leading to Strepsiptera and Diptera are long enough to be attracted in a parsimony analysis even if they were separated in the true phylogeny. I simulated data sets in which Strepsiptera were placed on every possible branch of the tree estimated by Carmean and Crespi (Fig. 2). For each of the 21 possible placements of Strepsiptera, 100 data sets were simulated. Branch lengths and other parameters (transition:transversion parameter, shape parameter of the gamma distribution, equilibrium nucleotide frequencies; see Table 1 for a description of the models used in this study) were estimated for each possible placement using the maximum likelihood method as implemented in PAUP* 4.0d52 (Swofford, 1996) and PAML

TABLE 1. Summary of the models of DNA substitution used in this study. The parameters of the models include the equilibrium nucleotide frequency (π), the transition:transverion rate bias (κ), and the shape parameter of the gamma distribution (α). When $\kappa = 1.0$, there is no transition:transversion bias. The discrete gamma distribution (with x rate categories) is used to model among-site rate variation. When $\alpha = \infty$, the rates at all sites are equal. "E" denotes free parameters that are estimated using maximum likelihood.

Model	Nucleotide frequencies				κ	α	Reference
	π_A	π_C	π_G	π_T			
JC69	0.25	0.25	0.25	0.25	1.0	∞	Jukes and Cantor (1969)
K80	0.25	0.25	0.25	0.25	E	∞	Kimura (1980)
K80+Γ_x	0.25	0.25	0.25	0.25	E	E	Kimura (1980); Yang (1993)
F81	E	E	E	E	1.0	∞	Felsenstein (1981)
HKY85	E	E	E	E	E	∞	Hasegawa et al. (1985)
HKY85+Γ_x	E	E	E	E	E	E	Hasegawa et al. (1985); Yang (1993)

(Yang, 1995), and these parameter estimates were used in the simulations. Data sets of the same size as the original (770 nucleotide sites) were simulated under the HKY85+Γ_∞ model of DNA substitution (Hasegawa et al., 1985; Yang, 1993, 1994) on each model tree and analyzed using maximum parsimony. For four of the model trees on which data were simulated, Strepsiptera were five branches away from the long branch leading to Diptera. However, parsimony analysis of these simulated data sets resulted in a Strepsiptera–Diptera relationship a high proportion of the time (0.53–1.00) regardless of where Strepsiptera were placed on the simulated trees. It appears that the branches leading to Diptera and Strepsiptera are long enough to attract each other in parsimony analysis.

If the Strepsiptera–Diptera relationship is the spurious result of long-branch attraction, then a method that corrects for the multiple substitutions that occur along long branches should provide an estimate in which the long branches are placed apart. Figure 1b shows the tree estimated by maximum likelihood implemented with a HKY85+Γ_5 model of DNA substitution (this model corrects for multiple substitutions and allows for rate heterogeneity among sites, different equilibrium nucleotide frequencies, and a transition : transversion rate bias; Hasegawa et al., 1985; Yang, 1993, 1994). The parameters of the HKY85+Γ_5 model (κ and α) were estimated for each tree using the likelihood criterion (using PAUP* 4.0d52; Swofford, 1996). Strepsiptera are not placed as the sister taxon to Diptera in the maximum likelihood tree (Fig. 1b). The phylogeny estimated using the maximum likelihood method not only places the long branches leading to Diptera and Strepsiptera in disparate parts of the tree, but it is consistent with some morphological evidence; Strepsiptera are placed as the sister taxon of the beetles (the meloid and *Tenebrio*), a relationship supported by morphological features such as posteromotorism (flight using metathoracic wings only; Kristensen, 1991). However, the bootstrap support for the Strepsiptera–meliod–*Tenebrio* clade is low (46%). Furthermore, the best tree under the constraint of Strepsiptera + *Drosophila* + *Aedes* monophyly is not significantly different from the maximum likelihood tree ($P = 0.54$; Kishino and Hasegawa, 1989). The tree estimated using maximum likelihood is insensitive to the model of DNA substitution assumed. The same tree is obtained if the method assumes the Jukes–Cantor (1969; JC69, log $L = -2965.93$), Kimura (1980; K80, log $L = -2916.39$, $\kappa = 2.97$), Felsenstein (1981; F81, log $L = -2959.07$), or Hasegawa et al. (1985; HKY85, log $L = -2909.77$, $\kappa = 2.92$) models. However, the HKY85+Γ_5 model represents the best fitting model; likelihood ratio tests indicate that adding parameters that account for unequal base composition (π_A, π_C, π_G, π_T), a transition: transversion bias (κ), and gamma distributed rate variation (α) provides a significant improvement in the likelihood score (Goldman, 1993).

Similar results are obtained if, instead of maximum likelihood, the minimum evolution method (Kidd and Sgaramella-Zonta, 1971) is used to estimate phylogeny. The implementation of the minimum evolution method in PAUP* 4.0d52 (Swofford, 1996) was used, and either simple proportion difference distances (p distances) or HKY85+Γ_5 distances were assumed. Figure 3 shows the trees estimated assuming different distance metrics. Figure 3a shows the tree estimated assuming p distances. This tree places Strepsiptera and Diptera as sister taxa. When the observed proportion differences between taxa are corrected for multiple substitutions, however, two sets of results are obtained. In some cases, trees in which Strepsiptera and Diptera form a monophyletic group are estimated (Fig. 3b), whereas in other cases, Strepsiptera and Diptera fall out in disparate parts of the tree (Fig. 3c). Whether a tree consistent with Strepsiptera + Diptera monophyly is estimated depends critically on whether among-site rate heterogeneity is accommodated. When among-site rate variation is not accounted for, a tree with Strepsiptera + Diptera as a monophyletic

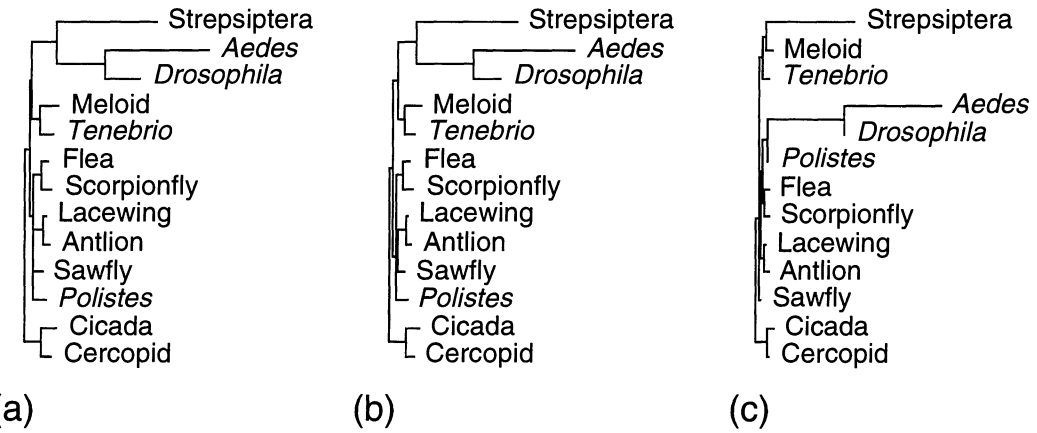

FIGURE 3. The trees estimated using the minimum evolution method differ depending on the distance metric used. (a) Tree estimated assuming p distances (tree length = 0.424 expected substitutions per site). (b) Tree estimated assuming HKY85 distance with equal rates for all sites ($\alpha = \infty$; tree length = 0.467 expected substitutions per site). (c) Tree estimated assuming HKY85+Γ_5 distances with $\alpha = 0.2$ (tree length = 0.720 expected substitutions per site).

group is estimated. However, if among-site rate variation is modeled using a gamma distribution and $0.125 < \alpha < 0.300$, then a tree placing Strepsiptera with the beetles is estimated. The maximum likelihood estimate of the gamma shape parameter is $\alpha = 0.29$, suggesting that a gamma shape parameter value consistent with Strepsiptera + beetle monophyly is best. Not only does the correction for multiple substitutions affect phylogeny estimation, but the type of correction that is applied also is important.

DISCUSSION

Although the results of this study are provocative, it is unknown how robust the result is to addition of more taxa and characters. The addition of more taxa to the branches leading to Diptera and to Strepsiptera would be expected to lessen the effect of the long branches in a parsimony analysis (Huelsenbeck, 1991; Hillis, 1996), but only if the additional branches are connected to the tree in an optimal way. If, for example, additional strepsipteran taxa all join near the tip of the existing strepsipteran taxon, then long-branch attraction can still be a problem in a parsimony analysis. It is also unknown whether this result is robust to alignment. The 18S rDNA sequence for Strepsiptera is very unusual in that it is very long (3,316 bp; Chalwatzis et al., 1995). Problems with alignment, however, were minimized by excluding sites with gaps or missing data.

Ribosomal sequences are very popular as phylogenetic markers. Thus, it is interesting, and perhaps informative, that 18S rDNA sequences provide ambiguous results for this insect phylogeny. The 18S rDNA sequence data also provided ambiguous results for the phylogeny of amniotes (mammals, birds, lizards, crocodiles, and turtles; Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996); for amniotes, the 18S sequences provided a phylogeny that contradicted the phylogeny from five other genes, and the incongruence could not be attributed to sampling error. Although two examples of anomalous results using 18S rDNA does not necessarily indicate that the gene is a poor phylogenetic marker, these results are of concern and should motivate a more general survey of phylogenies estimated using this gene.

This result has important ramifications for phylogenetic studies because it suggests that long-branch attraction is a real phenomenon, not just a theoretical one. The criteria outlined here should allow systematists to identify those cases in

which parsimony (or other methods) can be expected to have problems with long branches. However, a safer course may be to use methods that are less sensitive to the long-branch problem in the first place, especially if application of the tests proposed here indicates that long-branch attraction is a prevalent problem in nature.

ACKNOWLEDGMENTS

This work was supported by a Miller Fellowship awarded to J.P.H. David Hillis, Jim McGuire, Sharon Messenger, Axel Meyer, and Ziheng Yang provided useful comments on an earlier version of this manuscript. Brent Mishler made useful suggestions for additional analyses. Bernard Crespi and David Carmean kindly provided the 18S rDNA sequences they had meticulously collected.

REFERENCES

ARNETT, R. H. 1968. The beetles of the United States. A manual for identification. American Entomology Institute, Ann Arbor, Michigan.

CARMEAN, D., AND B. CRESPI. 1995. Do long branches attract flies? Nature 373:666.

CHALWATZIS, N., A. BAUR, E. STETZER, R. KINZELBACH, AND F. K. ZIMMERMAN. 1995. Strongly expanded 18S rRNA genes correlated with a peculiar morphology in the insect order Strepsiptera. Zoology 98:115–126.

CROWSON, R. A. 1960. The phylogeny of the Coleoptera. Annu. Rev. Entomol. 5:111–134.

CROWSON, R. A. 1981. The biology of the Coleoptera. Academic Press, London.

EFRON, B. 1979. Bootstrap methods: Another look at the jackknife. Ann. Stat. 7:1–26.

EFRON, B. 1982. The jackknife, the bootstrap, and other resampling plans. CBMS-NSF Reg. Conf. Ser. Appl. Math. No. 38. Society for Industrial and Applied Mathematics, Philadelphia.

FAITH, D. 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. 40:366–375.

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

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

FITCH, W. M. 1971. Toward defining the course of evolution: Minimal change for a specific tree topology. Syst. Zool. 20:406–416.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

HENDY, M. D., AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature 383:130–131.

HUELSENBECK, J. P. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Zool. 40: 458–469.

HUELSENBECK, J. P., AND J. J. BULL. 1996. A likelihood ratio test for detection of conflicting phylogenetic signal. Syst. Biol. 45:92–98.

HUELSENBECK, J. P., J. J. BULL, AND C. W. CUNNINGHAM. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. Munro, ed.). Academic Press, New York.

KATHIRITHAMBY, J. 1991. Strepsiptera. Pages 684–695 *in* The insects of Australia, 2nd edition (I. D. Naumann, P. B. Carne, J. F. Lawrence, E. S. Nielsen, J. P. Spradberry, R. W. Taylor, M. J. Whitten, and M. J. Littlejohn, eds.). CSIRO, Melbourne Univ. Press, Melbourne.

KIDD, K. K., AND L. A. SGARAMELLA-ZONTA. 1971. Phylogenetic analysis: Concepts and methods. Am. J. Hum. Genet. 23:235–252.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29: 170–179.

KRISTENSEN, N. P. 1991. Phylogeny of extant hexapods. Pages 125–140 *in* The insects of Australia, 2nd edition (I. D. Naumann, P. B. Carne, J. F. Lawrence, E. S. Nielsen, J. P. Spradberry, R. W. Taylor, M. J. Whitten, and M. J. Littlejohn, eds.). CSIRO, Melbourne Univ. Press, Melbourne.

ROSS, H. H., C. A. ROSS, AND J. R. P. ROSS. 1982. A text book of entomology, 4th edition. Wiley, New York.

SWOFFORD, D. L. 1996. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0. Sinauer, Sunderland, Massachusetts.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39: 306–314.

YANG, Z. 1995. Phylogenetic analysis using maximum likelihood (PAML). Institute of Molecular Evolution and Genetics, Pennsylvania State Univ., University Park.

Received 28 December 1995; accepted 11 September 1996
Associate Editor: Brian Farrell