Syst. Biol. 49(2):225–232, 2000

Invariable Sites Models and Their Use in Phylogeny Reconstruction

MIKE STEEL,[1] DANIEL HUSON,[2] AND PETER J. LOCKHART[3]

[1]*Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand;*
E-mail: m.steel@math.canterbury.ac.nz
[2]*Applied and Computational Mathematics,Fine Hall, Princeton University, Princeton, New Jersey*
08544–100, USA; E-mail: huson@math.Princeton.edu
[3]*Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand;*
E-mail: p.j.lockhart@massey.ac.nz

Abstract.—Phylogenetic inference is well known to be problematic if both long and short branches occur together in the underlying tree. With biological data, correcting for this problem may require simultaneous consideration for both substitution biases and rate heterogeneity between lineages and across sequence positions. A particular form of the latter is the presence of invariable sites, which are well known to mislead estimation of genetic divergences. Here we describe a capture-recapture method to estimate the proportion of invariable sites in an alignment of amino acids or nucleotides. We use it to investigate phylogenetic signals in 18S ribosomal DNA sequences from Holometabolus insects. Our results suggest that, as taxa diverged, their 18S rDNA sequences have altered in both their distribution of sites that can vary as well as in their base compositions. {Covariotide evolution; invariable sites; LogDet; sites free to vary.}

Given differences in the branch lengths of the true underlying evolutionary tree, uncorrected multiple substitutions can cause error in the reconstruction of both tree topology and branch lengths (Felsenstein, 1978; Hendy and Penny, 1989; Kim, 1996). This will cause problems for parsimony and incompatibility methods, as well as methods that use uncorrected distances. Model violation can also cause more statistically sophisticated methods, such as maximum likelihood, to be inconsistent. One example includes nonstationarity (e.g., Hasegawa and Hashimoto, 1993; Steel et al., 1993; Lockhart et al., 1994; Pesole et al., 1995; Jermin et al., 1996), in which the substitution model varies across the tree (as evidenced by variation in nucleotide frequencies) but is not part of the likelihood model. A second example occurs (even for stationary models) when positional rate heterogeneity (e.g., Yang, 1996; Van de Peer et al., 1996; Waddell et al., 1997) is not correctly incorporated in the analysis. A third complicating factor, which may also lead to inconsistency if not explicitly handled, is the presence of covarion/covariotide structure in the sequences (Fitch and Markowitz, 1970; Miyamoto and Fitch, 1995; Shoemaker and Fitch, 1989). In this case, tree building can be misled if distantly related sequences accepting parallel mutations also share similar distributions of variable sites (Lockhart et al., 1998; Philippe and Laurent, 1998).

Dealing with different causes of tree building inconsistency at the same time can be problematic, particularly for biological data (e.g., as discussed by Whitfield and Cameron, 1998). One approach is to use "invariable sites models" (Churchill et al., 1992; Reeves, 1992; Swofford et al., 1996)—sequence substitution models whose corresponding transformations are applied to sequence data after the removal of unvaried positions in an alignment of sequences. This approach follows from the observation that positional rate heterogeneity in sequences is sometimes well approximated by assuming that a certain proportion of sites (a subset of those that are constant or unvaried) are "invariable" (cannot change) and that the remaining sites evolve at a constant rate (Adachi and Hasegawa, 1995; Waddell, 1996). Invariable sites models can be implemented with correction formulae such as the LogDet /paralinear transformation (Steel 1994; Lake, 1994; Lockhart et al., 1994). The advantage of this transformation over stationary substitution models is that it allows the estimation of additive path lengths when taxa differ markedly in their base or amino acid compositions and when sequences show the type of positional rate heterogeneity that can be modeled by a mixture of invariable sites and constant rate sites.

In carrying out LogDet corrections, it is first helpful to estimate the proportion of sites in the alignment that are invariable. Maximum likelihood techniques can be used for such estimations (e.g., as implemented in PAUP*; Swofford, 1999). They require the use of a prespecified tree, and although these estimates are dependent on tree topology, deviations from the true underlying tree need to be significant to influence invariable sites estimates (Sullivan et al., 1996; Lockhart et al., 1998). By using a capture–recapture technique, a tree-independent method has been described by Sidow et al. (1992); however, this approach is inapplicable for noncoding DNA. Here we describe one further new approach that is also tree independent and can be computed efficiently for large numbers of coding or noncoding sequences. We use it with the LogDet correction to investigate an 18S ribosomal DNA data set previously noted for positional rate heterogeneity and compositional bias (Huelsenbeck, 1998).

The Fly Trap ("Strepsiptera Problem") Example

Uncertainty has attended the question of whether insect 18S and 28S sequences support Strepsiptera as sister taxa of Diptera (Whiting et al., 1997), or whether the joining of these taxa in some analyses is possibly an artifact of tree building, resulting from some form of long branch attraction (Huelsenbeck, 1997, 1998). Much of the discussion has concerned whether the branches leading to these taxa are sufficiently long enough to attract each other. Except for the study by Friedrich and Tautz (1997), however, little discussion has yet developed the issue of whether processes of substitution are sufficiently uniform across the true underlying tree to allow recovery of the correct phylogeny. We make observations on the 18S rDNA sequences that highlight the complexity of the substitution patterns in Holometabolus insects. Our study raises issues we believe need to be further explored, given their likely importance for understanding both the nature of the present Strepsiptera controversy and the implementation of invariable sites models in sequence analyses.

Figure 1 shows the unusual phylogenetic structure, represented by Split Decomposition (Bandelt and Dress, 1992; Huson, 1998), of the 18S rDNA data studied by Huelsenbeck (1997). In this tree, the long branch taxa join as sister species, and the question asked is whether such a relationship represents the true phylogeny or whether this sister relationship is an artifact of long branch attraction (Felsenstein, 1978; Hendy and Penny, 1989; Kim, 1996). To investigate this, we examined support for two groupings, (Strepsiptera, *Aedes, Drosophila*) and (Strepsiptera, Meloid, *Tenebrio*), under LogDet/distance Hadamard (Penny et al., 1996), LogDet/minimum evolution, and LogDet/neighbor joining (e.g., Saitou and Imanishi, 1989). The first grouping is

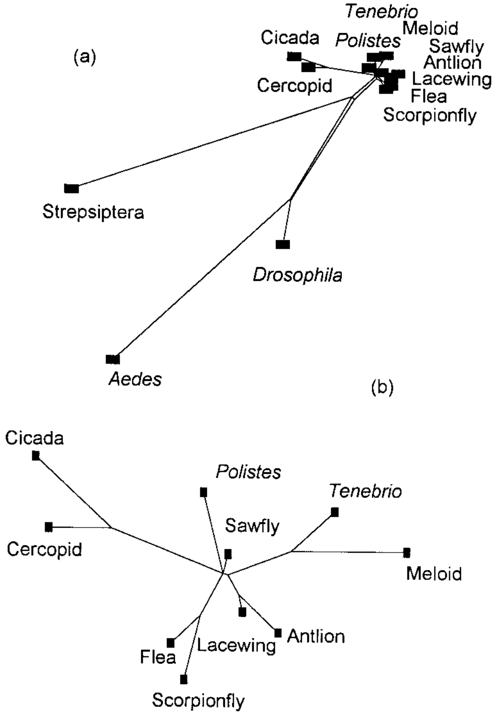

Figure 1. Splitsgraph with Hamming distances (number of observed differences/sequence length) calculated by using all sequence positions (768 unambiguous sites with no gaps or missing data). (a) Thirteen insect taxa. Long branches lead to Strepsiptera, *Aedes*, and *Drosophila*. (b) Same as (a) but with Strepsiptera, *Aedes*, and *Drosophila* removed. In a splitsgraph the strongest signals in the data that do not fit onto a unique bifurcating tree appear as reticulations.

the one favored in tree reconstructions that assume all sites in the data are equally variable (e.g., as done in Fig. 1); the second grouping is the one favored by some analyses that assume some sites in the data are invariable or are changing at different rates (Huelsenbeck, 1997). We use the LogDet correction (Lockhart et al., 1994) because the 18S rDNA sequences in this data set are characterized by compositional biases (Table 1).

For these data we have obtained capture–recapture (using Splitstree 3; Huson, 1998) and HKY85 maximum likelihood estimates (using PAUP*4; Swofford, 1999) for the proportion of invariable sites. We use these estimates to help provide a framework for interpreting the distance Hadamard spectra and the tree-building results shown in Figures 1 and 2. We also use the capture–recapture method in implementing an inequality test (Lockhart et al., 1998) for determining whether two groups of sequences may differ in their distribution of variable sites. The capture–recapture method differs from the approach of Sidow et al. (1992) in that, instead of performing capture–recapture on pairs of sites (in a codon), we consider pairs of sequences.

CAPTURE–RECAPTURE ESTIMATES OF INVARIABLE SITES

Suppose sequences evolve under a model in which a certain proportion ν of sites are variable and the remaining sites are invariable (unable to undergo substitution). We assume that the variable sites (some of which may be constant across the species, and thus be indistinguishable from unvaried sites) evolve independently and identically according to a Markov model on the underlying evolutionary tree. Such a model assumes that all variable sites evolve at the same rate (however, if rates of change differ amongst the variable sites, the estimates below still provide a lower bound to the number of variable sites.)

Let f_{ij} be the proportion of sites in which sequence i is in a different state to sequence j. For four distinct sequences i, j, k, l, let $f_{ij|kl}$ denote the proportion of sites where sequence i is in a different state to sequence j, and sequence k is in a different state to sequence l. Suppose for the moment that we know that the underlying evolutionary tree separates the pair of sequences i, j from the pair of sequences k, l by at least one edge. We first show that we may estimate ν by $\frac{f_{ij}f_{kl}}{f_{ij|kl}}$, whereas the other two ratios, $\frac{f_{ik}f_{jl}}{f_{ik|jl}}$ and $\frac{f_{il}f_{jk}}{f_{il|jk}}$, tend to underestimate ν. The justification of these statements is as follows: For r,s in $[i, j, k, l]$, let D_{rs} denote the event that, at a random site, the sequences r and s are in different states. Let V denote the event that the random site is variable. Then, under

TABLE 1. A, G, C, T content at 110 parsimony sites in the 18S rDNA alignment of insect taxa.

Taxon	A	C	G	T
Strepsiptera	0.273	0.255	0.191	0.282
Aedes	0.327	0.227	0.182	0.264
Drosophila	0.318	0.173	0.164	0.345
Flea	0.164	0.309	0.264	0.264
Scorpionfly	0.191	0.282	0.236	0.291
Lacewing	0.182	0.291	0.273	0.255
Antlion	0.182	0.273	0.273	0.273
Sawfly	0.173	0.309	0.273	0.245
Meloid	0.209	0.300	0.255	0.236
Polistes	0.227	0.264	0.236	0.273
Tenebrio	0.191	0.373	0.273	0.164
Cicada	0.155	0.355	0.300	0.191
Cercopid	0.145	0.336	0.291	0.227
Mean	0.210	0.288	0.247	0.255

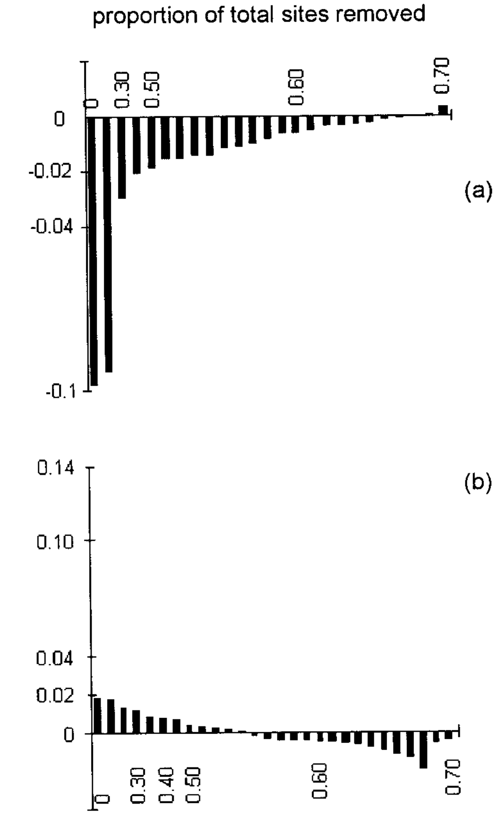

FIGURE 2. Distance Hadamard spectra showing (a) net support for various partitions of the Strepsiptera, *Meloid, Tenebrio* partition (frequency of patterns supporting minus frequency of patterns contradicting). (b) Net support for the Strepsiptera, *Aedes, Drosophila* partition. Strength of signals was calculated under the LogDet transformation as an increasing proportion of unvaried sites were removed. All unvaried sites are removed when 0.71 sites have been eliminated.

models such as the Kimura 3ST, we have, for the conditional probabilities:

$$P\{D_{ij}\&D_{kl} \mid V\} = P\{D_{ij} \mid V\}P\{D_{kl} \mid V\} \qquad (1)$$

and

$$P\{D_{ik}\&D_{jl} \mid V\} \geq P\{D_{ik} \mid V\}P\{D_{jl} \mid V\} \qquad (2)$$

$$P\{D_{il}\&D_{jk} \mid V\} \geq P\{D_{il} \mid V\}P\{D_{jk} \mid V\} \qquad (3)$$

Now, for two distinct sequences r,s [$[i, j, k, l]$

$$P\{D_{rs}\} = P\{D_{rs} \mid V\}P\{V\} \qquad (4)$$

because event D_{rs} can occur only at a variable site. Similarly,

$$P\{D_{ij}\&D_{kl}\} = P\{D_{ij}\&D_{kl} \mid V\}P\{V\}. \qquad (5)$$

Combining Equations 1, 4, and 5, we obtain

$$P\{V\} = \frac{P\{D_{ij}\}P\{D_{kl}\}}{P\{D_{ij}\&D_{kl}\}} \qquad (6)$$

Substituting the analogs of Equation 5 (for $D_{ik}\&D_{jl}$ and $D_{il}\&D_{jk}$) into Equations 2, 3, and 4, we obtain

$$P\{V\} \geq \frac{P\{D_{ik}\}P\{D_{jl}\}}{P\{D_{ik}\&D_{jl}\}}, \frac{P\{D_{il}\}P\{D_{jk}\}}{P\{D_{il}\&D_{jk}\}} \qquad (7)$$

If we now estimate $P\{D_{rs}\}$ by f_{rs}, $P\{D_{rs}\&D_{uv}\}$ by $f_{rs \mid uv}$ and v by $P\{V\}$, then from Equations 6 and 7 we deduce that:

$$v \approx \frac{f_{ij}f_{kl}}{f_{ij \mid kl}} \geq \frac{f_{ik}f_{jl}}{f_{ik \mid jl}}, \frac{f_{il}f_{jk}}{f_{il \mid jk}}$$

as claimed. Equivalently,

$$v \approx max \left[\frac{f_{ij}f_{kl}}{f_{ij \mid kl}}, \frac{f_{ik}f_{jl}}{f_{ik \mid jl}}, \frac{f_{il}f_{jk}}{f_{il \mid jk}}\right] \qquad (8)$$

A useful feature of this last equation is that, by symmetry, it remains true even if the underlying tree connecting the four leaves (i, j, k, l) is one of the other two trees, and thus the assumption we made earlier that we know how the underlying tree resolves the four sequences turns out to be unnecessary. Of course the proportion of variable sites lies between $1 - u_{ijkl}$ and 1, where u_{ijkl} is the proportion of sites that are observed to be constant (unvaried) across the four species. We may thus refine the estimate of v by taking

$$v = min \left\{1, max \left(1 - u_{ijkl}, \frac{f_{ij}f_{kl}}{f_{ij \mid kl}}, \frac{f_{ik}f_{jl}}{f_{ik \mid jl}}, \frac{f_{il}f_{jk}}{f_{il \mid jk}}\right)\right\}$$

So far we have considered only one quartet of sequences, so when there are more than four sequences, a more accurate estimate of v should be that given by averaging the estimate over all quartets. If the number of sequences is large (≥50 or more) this may take too long, so we instead average over a large randomly selected subset of quartets. In both cases it is useful to estimate the standard deviation of the quartet estimates of v to see how much these estimates are dispersed across different quartets. This should not be confused with the standard deviation estimate for a single quartet capture–recapture, which measures the variation resulting from the finite sequence length (to calculate this second estimate of standard deviation, see Thompson, 1992:214, eq. 4). Thus, the invariable sites estimates obtained for quartets have been averaged and are shown with their standard deviation in Table 2.

The justification of our approach relies on Equations 1–3, which are provably exact under certain models such as the Kimura 3ST model, though they are only approximations for more general models. Equation 1 was independently noted by Waddell (1996) as part of a more restrictive capture–recapture estimation technique (see also Waddell et al., 1999).

Spectral Plots for LogDet-Invariable Sites Model

Figure 2 shows a Hadamard/LogDet spectral plot (Lockhart et al., 1999) indicating support for the Strepsiptera, Meloid, *Tenebrio* (Figure 2a) and Strepsiptera, *Aedes, Drosophilia* (Figure 2b) partitions (groupings) as various sites that are unvaried in the alignment are successively removed. Stronger support for the Strepsiptera, Meloid, *Tenebrio* partition over the Strepsiptera, *Drosophila, Aedes* partition occurs only when most of the sites observed to be constant or unvaried in the data set (i.e., corresponding to 0.63–0.70 of the total sequence length) are removed from the analysis before LogDet paths are calculated. Under minimum evolution/LogDet, this first partition is selected when 0.59–0.68 sites are removed; under neighbor joining/LogDet, this partition is recovered when 0.59–0.70 sites are removed. The observation that support for the Strepsiptera, *Drosophila,* and *Aedes* partition decreases as unvaried sites are removed, whereas support for the Strepsiptera, Meloid and *Tenebrio* partition becomes favored, raises concern about the reliability of the Strepsiptera, *Drosophila,* and *Aedes* grouping. This is particularly the case since this switching of support occurs after we have removed that proportion of 18S rDNA sites estimated as being invariable, i.e., ∼>0.6 sites (Table 2).

Estimates of Invariable Sites in 18S rDNA Alignments

Using our capture–recapture procedure and a maximum likelihood procedure, we have made estimates of the number of sites that are invariable (shown in Table 2). They indicate that a significant proportion of the sites in these 18S ribosomal DNA data are invariable; moreover, for these data, the capture–recapture estimates are similar to those obtained by a maximum likelihood procedure. Nevertheless, despite similar estimates obtained, in general, such estimations cannot be used to precisely pinpoint the proportion of unvaried sites that should be excluded before tree building. The reason for this concerns not only statistical error associated with the estimation procedures, but also the nonuniform processes of evolution between sequences. For example, in the 18S rDNA data studied here, this is suggested from the observation that estimates of invariable sites appear to be greater in the 13 taxon data set than in the 11 taxon data set. We test this possibility of covariotide structure in the following section, using a inequality test described recently in Lockhart et al. (1998).

Note that if invariable sites estimates are made when sequences show covarion or covariotide structure, the values obtained from our quartet procedure and those from maximum likelihood procedure may well

Table 2. Estimates of the proportion of the total number of sequence sites that are invariable made by using our capture–recapture method and the HKY85 maximum likelihood model (Swofford et al., 1996).

Method	All taxa ($N = 13$)	Beetles and Strepsiptera removed ($N = 10$)
Capture–recapture	0.631 + / −0.176[a]	0.734 + / −0.220[a]
HKY ML/B&B parsimony tree(s)	0.618[b]	0.815

[a] Standard deviation between quartets.

[b] Mean of 27 trees.

differ. This will occur because some quartets used to estimate the proportion of variable sites with the capture–recapture procedure will be sampled not only across sequences with different distributions of variable and invariable sites but also within covarion/covariotide groups (i.e., between sequences having similar distributions of invariable sites). Thus, the average estimate of invariable sites between quartets may well exceed the maximum likelihood estimate depending on the taxon density within different covarion/covariotide groups. Similarly, covarion/covariotide structure will also be expected to increase the standard deviation associated with the mean average estimate for quartets.

Non-I.I.D. Sequence Evolution and Invariable Sites Estimates

Virtually all methods for tree building assume that sequence positions evolve identically and independently (i.i.d.). This is also assumed with our capture–recapture procedure for estimating invariable sites. However, with some biological sequences, sequence evolution is possibly more complex, with homologous sequences sometimes differing in their distributions of sites free to vary (a result of sequences having undergone asymmetric covarion/covariotide shifts). The results in Table 2 show that estimates of invariable sites are higher when Strepsiptera, Meloid, and *Tenebrio* are not included in the estimation, suggesting that these taxa may differ in their distribution of variable sites from those of the other insect sequences present in the data set we studied. This conclusion is supported by results of our covarion/covariotide inequality test (see Lockhart et al., 1998, for details). In this test, site patterns in an alignment of data are characterized into five pattern classes, and the test determines whether there is a disproportionally large number of N_3 and N_4 patterns for the sequences to have evolved under an i.i.d. model of evolution. With the present data, the relative numbers of the five observed classes are N_1 (538), N_2 (8), N_3 (31), N_4 (136), and N_5 (55). To make this test we need to estimate the proportion of observed N_1 patterns (unvaried sites) that are actually variable in our alignment. If we use our two capture–recapture estimates from Table 2, these suggest that either all unvaried sites (538) are invariable, i.e., $N_1 = 0$ (estimate made from using 10 taxa), or as many as 54 may be variable, i.e., $N_1 = 54$ (estimate made from using 13 taxa). Using these figures we infer a difference in the distribution of variable sites between *Aedes*, *Drosophila*, and Strepsiptera and the other 10 insects at the 0.05 ($Z = -5.82684$, $N_1 = 0$) and 0.1 levels (then $Z = -1.25124$; $N_1 = 54$) of significance, respectively.

Discussion

With biological sequences a number of causes of inconsistency may be involved in tree reconstruction (when using uncorrected distances/ parsimony or overly simplistic implementation of likelihood), beyond the interplay of long and short branches first described by Felsenstein (1978). Ignoring positional rate heterogeneity may cause inconsistency and make difficult the recognition of substitution biases, which leads to inconsistency of tree building methods. Asymmetric processes of substitution resulting in taxon differences in nucleotide/ amino acid compositions (via nonstationarity) and differences in the distributions of variable sites are also potential problems.

Tables 1 and 2 suggest that both types of asymmetric change (compositional bias and covariotide differences) characterize 18S rDNA sequences from some Holometabolus insects, and both phenomena may be contributing to both the observed branch length differences between taxa and the difficulty of correctly placing species in reconstructed phylogenetic trees. Nevertheless, demonstrating nonuniformity of substitution processes is not equivalent to demonstrating inconsistency in tree building, because one might reasonably argue that similar compositional biases and altered distributions of variable sites in Strepsiptera and Diptera are themselves characteristic of a close phylogenetic affinity. This question needs further investigation. Such a study should now be possible by using the larger 18S rDNA data set available (Whiting et al., 1997). Interestingly, in this larger data set, as in the smaller data set we study here, the branches leading to some of the more recently determined taxa also show extreme length differences. This may suggest an irregular pattern of nonuniform/asymmetric shifts in the evolution of 18S rDNA insect rDNA sequences. Such irregular covarion/covariotide changes in sequence evolution have been suggested elsewhere (Philippe et al., in press), and the analysis of more sequences (particularly if based on structural alignments) should allow determination of the extent and regularity of any such changes in evolutionary processes. Understanding the regularity of patterns of change should help in evaluating the reliability of trees reconstructed from insect rDNA.

If nucleotide substitution processes are complex, that is, asymmetric in respect of both composition and covariotide structure across the true underlying tree, it is unlikely that parsimony, simple substitution models, or their associated transformations will allow for the reliable reconstruction of the relationships between all taxa. In this case, testing the support for individual partitions with transformations, such as LogDet incorporating invariable sites elimination, could help provide one means for evaluating the robustness of different phylogenetic groupings.

Acknowledgments

We thank John Huelsenbeck and Michael Whiting for sending ribosomal DNA alignments, Karl Kjer for helpful comments, the New Zealand Marsden Fund, Alexander von-Humboldt Foundation, and NZ/FRG cooperation (S.T.C.) agreement program for supporting this research.

References

Adachi, J., and M. Hasegawa. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: Heterogeneity among amino acids. J. Mol. Evol. 40:622–628.

Bandelt, H. J., and A. W. M. Dress. 1992. Split decomposition: A new and useful approach to phylogenetic distance data. Mol. Phylogenet. Evol. 1:242– 252.

Churchill, G. A., A. von Haeseler, and W. C. Navidi. 1992. Sample size for a phylogenetic inference. Mol. Biol. Evol. 9:753–769.

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

Fitch, W. M., and E. Markowitz. 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.

Friedrich, M., and D. Tautz. 1997. An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. Mol. Biol. Evol. 14:644–653.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:310–321.

Huelsenbeck, J. P. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46:69–74.

Huelsenbeck, J. P. 1998. Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved? Syst. Biol. 47:519–537.

Huson, D. H. 1998. SplitsTree: A program for analyzing and visualizing evolutionary data. Bioinformatics 14:68–73.

Jermin, L. S., P. G. Foster, D. Graur, R. M. Lowe, and R. H. Crozier. 1996. Unbiased estimation of symmetrical directional mutation pressure from protein coding DNA. J. Mol. Evol. 42:476–480.

Kim, J. 1996. General inconsistency conditions for maximum parsiomony: Effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45: 363–374.

Lake, J. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA 91:1455–1459.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930–1934.

Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. Huson, M. A. Charleston, and C. J. Howe. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15:1183–1188.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12:503–513.

Penny, D., M. D. Hendy, P. J. Lockhart, and M. A. Steel. 1996. Corrected parsimony, minimum evolution, and Hadamard conjugations. Syst. Biol. 45: 596–606.

Pesole, G., G. Dellisanti, G. Preparata, and C. Saccone. 1995. The importance of base composition in the correct assessment of genetic distance. J. Mol. Evol. 41:1124–1127.

Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8: 616–623.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Müller, and H. Le Guyader. (in press) Tree reconstruction and the phylogeny of the eukaryotes. Proc. R. Soc. Lond. B.

Reeves, J. H. 1992. Heterogenetity in the substitution process of amino acid sites of proteins coded for by mitochonadrial DNA. J. Mol. Evol. 35:17–31.

Saitou, N., and T. Imanishi. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. 6:514–525.

SHOEMAKER, J. S., AND W. M. FITCH. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270–289.

SIDOW, A., T. NGUYEN, AND T. P. SPEED. 1992. Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. 35:253–260.

STEEL, M. A. 1994. Recovery of a tree from the leaf coloration it generates under a Markov model. Appl. Math. Lett. 7:19–23.

STEEL, M. A., P. J. LOCKHART, AND D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.

SULLIVAN, J., K. E. HOLSINGER, AND C. SIMON. 1996. The effect of topology on estimates of among site rate variation. J. Mol. Evol. 42:308–312.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic Inference. Pages 459–461 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinaur, Sunderland, Massuchusetts.

SWOFFORD, D. L. 1999. PAUP 4.65 Sinaur, Sunderland. Massuchusetts.

THOMPSON, S. K. 1992. Sampling. John Wiley and Sons, New York.

VAN DE PEER, Y., S. A. RENSING, U.-G. MAIER, AND R. DEWACHTER. 1996. Substitution rate calibration of small subunit ribosomal subunit RNA identifies Chlorarachnida nucleomorphs as remnants of green algae. Proc. Natl. Acad. Sci. USA 93:7732–7736.

WADDELL, P. J. 1996. Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood. Ph.D. dissertation, Massey Univ., Palmerston North, New Zealand.

WADDELL, P. J., D. PENNY, AND T. MOORE. 1997. Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol. Phylogenet. Evol. 8:33–50.

WADDELL, P. J., C. YING, J. HAUF, AND M. HASEGAWA. 1999. Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid–invariant sites-LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. Syst. Biol. 48:31–53.

WHITING, M. F., J. C. CARPENTER, Q. D. WHEELER, AND W. C. WHEELER. 1997. The Strepsiptera problem: Phylogeny of the Holmetabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst. Biol. 46:1–68.

WHITFIELD, J. B., AND S. A. CAMERON. 1998. Hierarchical analysis of variation in the mitochondrial 16S rRNA gene among Hymenoptera. Mol. Biol. Evol. 15:1728–1743.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.

Received 1 February 1999; accepted 17 May 1999
Associate Editor: C. Simon