

Points of View

Syst. Biol. 53(4):638–643, 2004
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490468648

The Biasing Effect of Compositional Heterogeneity on Phylogenetic Estimates May be Underestimated

LARS S. JERMIIN,^{1,2} SIMON Y. W. HO,^{1,4} FAISAL ABABNEH,³ JOHN ROBINSON,³ AND ANTHONY W. D. LARKUM^{1,2}

¹*School of Biological Sciences, University of Sydney, NSW 2006, Australia; E-mail: lsj@bio.usyd.edu.au (L.S.J.)*

²*Sydney University Biological Informatics and Technology Centre, University of Sydney, NSW 2006, Australia*

³*School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia*

The effect of compositional heterogeneity in sequence data on phylogenetic inference was first identified as a potential problem in the late 1980s and early 1990s (Chang and Campbell, 2000; Conant and Lewis, 2001; Foster and Hickey, 1999; Hasegawa et al., 1993; Klenk et al., 1994; Lockhart et al., 1992a, 1992b; Loomis and Smith, 1990; Olsen and Woese, 1993; Penny et al., 1990; Sogin et al., 1993; Tarrío et al., 2001; Van Den Bussche et al., 1998; Weisburg et al., 1989), and by 1993 the first methods had been developed to measure the extent of the problem (Lockhart et al., 1993, 1994; Steel et al., 1993, 1995) or to overcome it (Foster, 2004; Galtier and Gouy, 1995, 1998; Galtier et al., 1999; Gu and Li, 1996, 1998; Lake, 1994; Steel, 1994; Steel et al., 1993, 1995; Tamura and Kumar, 2002; Yang and Roberts, 1995). It is now widely accepted that compositional heterogeneity in aligned sequence data can mislead methods commonly used to infer phylogenetic trees, but it is still unclear (i) why phylogenetic studies based on the LogDet (or paralinear) distance (Lockhart et al., 1994; Steel, 1994) sometimes fail to recover the expected tree topology from compositionally heterogeneous alignments (e.g., Foster and Hickey, 1999; Tarrío et al., 2001), and (ii) how much compositional convergence is necessary before the phylogenetic methods fail to recover the correct topology. Using Monte Carlo simulations to address the second point, Conant and Lewis (2001) concluded that “rather extreme amounts of convergence are necessary before parsimony begins to prefer the incorrect tree.” Other simulation studies have reached similar conclusions (e.g., Galtier and Gouy, 1995; Rosenberg and Kumar, 2003; Van Den Bussche et al., 1998). Based on the study by Galtier and Gouy (1995), it would appear that it is safe to use DNA for phylogenetic inference as long as the difference in GC content is less than 8% to 10%. This im-

plication, however, may simply be the product of the relative length of the internal edge. Here we investigate this issue and show that compositional heterogeneity in sequence data increases the difficulty with which short internal edges can be inferred using the maximum-parsimony method, the maximum likelihood method with an F81 model of nucleotide substitution, and the neighbor-joining method with distances estimated using the Jukes-Cantor model of nucleotide substitutions. We also show that the neighbor-joining method, with distances estimated using the LogDet method, has no difficulty in inferring the internal edge under conditions where the three other methods failed, and we conclude that as the number of sequences in phylogenetic data increases, so does the potential problem caused by compositional heterogeneity, and the need to assess whether the assumption of compositional homogeneity is violated by the data intended for phylogenetic analysis.

MATERIALS AND METHODS

Sequence Data

We tested the ability of four phylogenetic methods to recover the correct topology by analyzing data sets generated by Monte Carlo simulation using the program Hetero (Jermiin et al., 2003). During these simulations, the diverging nucleotide sequences were allowed to become increasingly GC-rich along the edges *a* and *d*, and increasingly AT-rich along the edges *b* and *c* (Fig. 1). One thousand alignments of 10,000 nucleotides were generated on trees with a fixed root-to-tip distance (0.5 time units) but with different internal edge lengths ($e + f = 0.01, 0.025, \text{ and } 0.05$ time units). In other words, if the first divergence (i.e., AB|CD) took place 100 million years ago, then the study aims to determine how likely it is to infer this event when the two subsequent divergences (i.e., A|B and C|D) took place 1.0, 2.5, or 5.0 million years later.

⁴*Present address: Henry Wellcome Ancient Biomolecules Centre, Department of Zoology, University of Oxford, OX1 3PS, United Kingdom.*

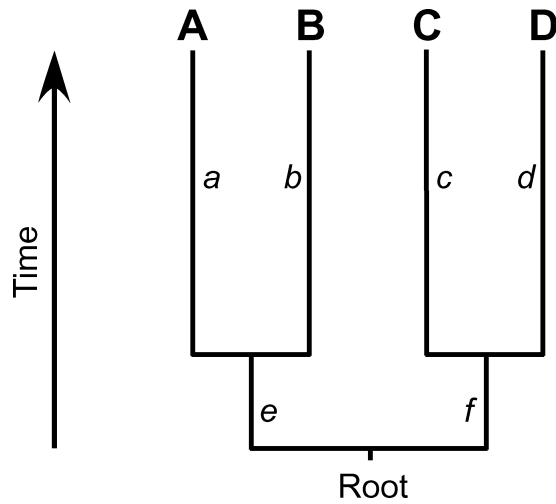


FIGURE 1. The phylogenetic tree used by Hetero. The root (Root), the edges (a, \dots, f) and the tips (A, B, C, D) are labeled, and the direction of time from the past to the present is indicated by an arrow. Each Monte Carlo simulation begins with a randomly generated ancestral nucleotide sequence at the root of the tree; this sequence is duplicated into two arrays. The sequences in these two arrays are allowed to evolve along the edges e and f . At a later stage during the simulation, each of the sequences in these two arrays is duplicated into another two arrays, and the final simulation occurs along the edges a, b, c , and d . The simulations along the individual edges are independent and may proceed by using different substitution models.

The rate of nucleotide substitution per site per time unit for the i th edge of the tree (Fig. 1) was given by

$$\mathbf{R}_i = \begin{bmatrix} -\sum_{y \neq A} \alpha_{iAy} \pi_{iy} & \alpha_{iAC} \pi_{iC} & \alpha_{iAG} \pi_{iG} & \alpha_{iAT} \pi_{iT} \\ \alpha_{iCA} \pi_{iA} & -\sum_{y \neq C} \alpha_{iCy} \pi_{iy} & \alpha_{iCG} \pi_{iG} & \alpha_{iCT} \pi_{iT} \\ \alpha_{iGA} \pi_{iA} & \alpha_{iGC} \pi_{iC} & -\sum_{y \neq G} \alpha_{iGy} \pi_{iy} & \alpha_{iGT} \pi_{iT} \\ \alpha_{iTA} \pi_{iA} & \alpha_{iTC} \pi_{iC} & \alpha_{iTG} \pi_{iG} & -\sum_{y \neq T} \alpha_{iTy} \pi_{iy} \end{bmatrix} \quad (1)$$

where α_{ixy} is the conditional rate of change from nucleotide x to nucleotide y in \mathbf{R}_i , and π_{iy} is the frequency of nucleotide y in \mathbf{R}_i . In each simulation we used $\alpha_{ixy} = 0.4$, $\pi_{ey} = \pi_{fy} = 0.25$, and $\pi_{Oy} = 0.25$ (π_{Oy} is the frequency of nucleotide y in the ancestral sequence); all the remaining parameters (i.e., π_{ay} , π_{by} , π_{cy} , and π_{dy}) were allowed to vary in order to generate the intended differences in the nucleotide content.

Phylogenetic Analyses

Phylogenetic trees were inferred from the simulated data using the following methods from the PHYLIP program package (Felsenstein, 2002): (i) the maximum-parsimony method (Fitch, 1971); (ii) the maximum-likelihood method with an F81 model of nucleotide substitution (Felsenstein, 1981); (iii) the neighbor-joining method (Saitou and Nei, 1987) with JC-corrected distances (Jukes and Cantor, 1969); and (iv) the neighbor-

joining method with LogDet distances (Lockhart et al., 1994). When alternative options were available during the phylogenetic analyses, we allowed all sites to evolve at the same rate and used a transition-transversion ratio of 0.5. In so doing, we analyzed the data in a manner that is consistent with the evolutionary processes that gave rise to these data; it also ensures that our results will not be confounded by the other factors that are known to bias phylogenetic estimates, such as among-site rate variation (Yang, 1996a).

RESULTS AND DISCUSSION

Phylogenetic Results Inferred Assuming Compositional Homogeneity

When the internal edge length was 0.05, the maximum-parsimony method recovered the true tree topology every time even though some compositional differences existed in the sequence data (Fig. 2). However, when the difference in the GC content exceeded $\sim 9\%$, the frequency of recovering the correct tree topology dropped sharply and reached 0% when the difference in the GC content exceeded $\sim 15\%$. When the internal edge length was halved to 0.025, the same pattern of successful phylogenetic recovery was observed but this time the transition from 100% to 0% occurred between $\sim 4\%$ and $\sim 12\%$ difference in the GC content. In other words, the length of the internal edge determined the probability of recovering the correct tree topology when compositional differences were present in the data. Reducing the internal edge length further to 0.01 produced results corroborating this observation, but they also showed that the maximum-parsimony method was not always able to infer the correct topology, even though the compositional differences might be extremely small. This last result is consistent with our expectations.

When the maximum-likelihood method was used to analyze the data, similar results were obtained, but there was a tendency for the maximum-likelihood method to recover the correct topology more frequently than the maximum-parsimony method, especially when a large compositional difference was present in the data (Fig. 2). Similar results were obtained when the neighbor-joining method was used to analyze matrices of JC-corrected distances. Notwithstanding the small differences in performance of these phylogenetic methods, it is clear that the probability of recovering a correct internal edge of a given length depends on the size of compositional difference in the sequence data used to estimate the phylogeny; shorter internal edges are more difficult to infer from sequence data than the longer internal edges, particularly when there is compositional heterogeneity in the data.

The results presented above show that three of the most commonly used phylogenetic methods are unable to recover the correct tree topology when compositional heterogeneity is present in the phylogenetic data—in so doing, they corroborate earlier results (e.g., Galtier and Gouy, 1995). However, the results outlined above also show that the chance of recovering a correct tree topology, from a data set with a given compositional

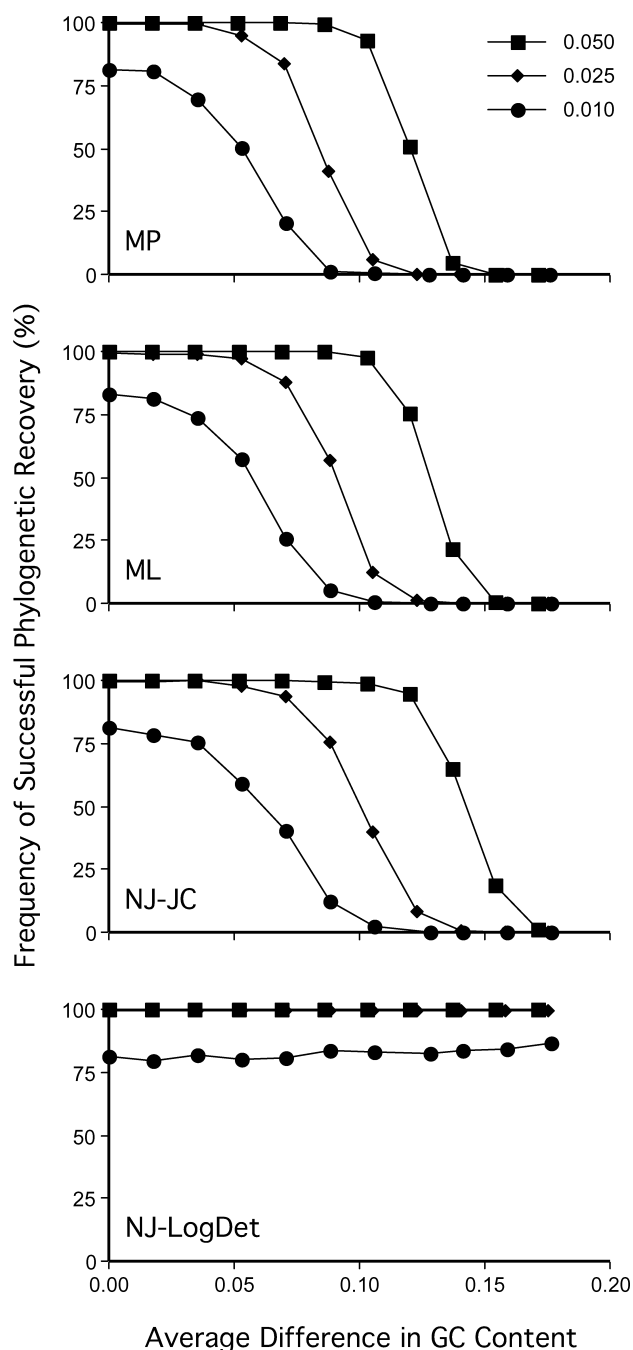


FIGURE 2. Results showing the rate of recovery of the correct phylogenetic tree as a function of phylogenetic method; the true length of the correct internal edge; and the observed average difference in GC content between A and D versus B and C. The legend in the top right corner refer to the length of the internal edge. Abbreviations: MP = maximum-parsimony method; ML = maximum-likelihood method; NJ-JC = neighbor-joining method (distances corrected using the Jukes-Cantor method); NJ-LogDet = neighbor-joining method (distances based on the LogDet method).

difference, decreased with the length of the correct tree's internal edge, which, although perhaps being obvious in hindsight, is still an important finding because it counters an emerging trend of considering compositional heterogeneity as a minor problem in phylogenetic inference

(Rosenberg and Kumar, 2003; Van Den Bussche et al., 1998). Interestingly, the results presented above corroborate the observation made by Conant and Lewis (2001) that "convergence in nucleotide composition is exacerbated by small internal branch lengths." However, Conant and Lewis did not elaborate on the important implications of their observation—instead, they concluded that "extreme combinations of substitution rates, transition/transversion bias, and equilibrium frequencies are required before parsimony is expected to fail."

Phylogenetic Results Inferred Assuming a General Model of Nucleotide Substitution

When the matrices of LogDet distances were used to infer neighbor-joining trees, the results were remarkably different—compositional heterogeneity did not affect the performance of this phylogenetic method; it recovered the correct topology nearly every time and it was only when the true internal edge was short that the frequency of successful phylogenetic recovery fell below 100% (Fig. 2). This observation is particularly interesting because several studies of empirical data have led to concern about the conditions under which the LogDet method will deal correctly with compositional heterogeneity (e.g., Chang and Campbell, 2000; Conant and Lewis, 2001; Foster and Hickey, 1999; Tarrío et al., 2001); indeed, Tarrío et al. (2001) concluded that the LogDet method "cannot completely overcome the distorting effects of the compositional variation that exists among the species of the *Drosophilidae*," whereas other authors suggested that other factors, such as rate heterogeneity among lineages (Foster and Hickey, 1999) and among sites (Conant and Lewis, 2001), might have biased the phylogenetic estimate. Lockhart et al. (1998) and Steel et al. (2000) have also noted the potential problem and added complexity caused by evolving distributions and proportions of variable sites; this should be investigated. A simpler explanation is that some of the internal edges in the true tree, in fact, may be very short. Only Conant and Lewis (2001) showed their trees with inferred edge lengths included, and the lengths of these internal edges were indeed very short in each case; clearly, it cannot be ruled out that short internal edges in the true tree have added an additional degree of complexity to the above-mentioned studies by Foster and Hickey (1999), Chang et al. (2000), Conant and Lewis (2001), and Tarrío et al. (2001).

Implications and Consequences of the Results

The results outlined above are important because the edge lengths chosen for the tree used to generate the data are similar to those inferred in many phylogenetic studies (see e.g., Conant and Lewis, 2001) and because the compositional differences generated by Monte Carlo simulation in the present study are found in many phylogenetic data sets, particularly those derived from the metazoan mitochondrial genome (see e.g., Jermini et al., 1994). Hence, we can assume that our conclusions are relevant to empirical studies.

Given an outgroup, the lengths of the true internal edges will become smaller (on the average) when the number of ingroup taxa in the phylogenetic data increases. Consequently, it will become more difficult to estimate the internal edges, particularly when compositional heterogeneity exists in phylogenetic data, as shown by the results outlined here. Our study therefore sends a much more cautionary signal than those based on previous simulation studies (see e.g., Conant and Lewis, 2001; Galtier and Gouy, 1995; Van Den Bussche et al., 1998).

The results outlined above highlight the advantage and disadvantage of the LogDet method; i.e., that it is more likely than the other three phylogenetic methods to recover the true topology from compositionally heterogeneous alignments of sites that have evolved at the same rate, and that the probability of inferring the true topology using this method, like the other phylogenetic methods, is less than 100% when internal edges in the true tree are very short. This observation is consistent with research by Baake and von Haeseler (1999), which shows that the Logdet distance is tree-additive, and so will infer the right topology under the Markov model.

One implication of the results presented above is that phylogenetic research may face increasingly large problems when the number of sequences in phylogenetic data is increased and there is compositional heterogeneity in the data. One way to tackle this problem would be to minimize the number of taxa in these data but this is unlikely to happen, because there is a justified, omnipresent tendency to study increasingly large data sets. The problem must be solved in a different manner, if possible, because by reducing the number of sequences we may lose important information with bearings on, for example, the modelling of among-site rate variation and invariable sites.

Another implication of the results presented above is that there is a need for more flexible methods to infer trees from compositionally heterogeneous data. Tentative measures to develop such methods have already been taken (see e.g., Foster, 2004; Galtier and Gouy, 1995, 1998; Galtier et al., 1999; Gu and Li, 1996, 1998; Lake, 1994; Lockhart et al., 1994; Steel, 1994; Steel et al., 1993, 1995; Tamura and Kumar, 2002; Yang and Roberts, 1995) but, because phylogenetic trees increasingly are inferred from concatenated alignments of genes or proteins (see, e.g., Waddell et al., 1999), there is an increasing need for phylogenetic methods and programs that allow a user to consider gene- and site-specific differences in the nucleotide (or amino acid) content, the rates of change, and the distribution and proportion of invariable sites. Methods that go some way towards providing this degree of flexibility are still relatively rare but some have been implemented in the following phylogenetic programs: PAML (Yang, 1996b), PAUP* (Swofford, 2001), MrBayes (Huelsenbeck and Ronquist, 2001), and p4 (Foster, 2004).

Methods for Assessing Compositional Heterogeneity

Our results highlight the importance of assessing the compositional heterogeneity of phylogenetic data as the

number of sequences increases. There are several methods to assess compositional homogeneity in sequence data—they fall into four categories.

1. Methods belonging to the first category use Monte Carlo simulations or the multinomial distribution to obtain sequence-specific estimates of the standard deviation of the mean nucleotide or amino acid content; these estimates are then compared using graphs or tables (Lanave et al., 1984, 1986). Although these methods have some appeal due to their simplicity, they are of limited value for surveys of big data sets, and more appropriate methods are now available (see below). Invariant sites can, and should be, removed before these methods are used; the biasing effect of these sites on estimates of the mean and standard deviation does not appear to have been realized by those who employed this method (e.g., Hashimoto et al., 1994, 1995; Saccone et al., 1989).
2. Methods belonging to the second category compare the nucleotide or amino acid content from different sequences using a homogeneity test of an $n \times m$ contingency table, where n is the number of sequences in the alignment and m is the size of the alphabet (usually, $m = 4$ and 20 for nucleotides and amino acids, respectively). Some of these methods compare all the sequences at once (e.g., von Haeseler et al., 1993), whereas other methods compare all (e.g., Preparata and Saccone, 1987) or some (e.g., Downton and Austin, 1997) of the pairs of sequences. Swofford's (2001) program, PAUP*, implements this category of methods for $n \geq 2$ by allowing users to choose which sequences to include in the assessment of compositional homogeneity. Tree-Puzzle (Schmidt et al., 2002; Strimmer and von Haeseler, 1996) also implements this category of methods but unlike PAUP*, it compares the composition of each sequence to the unweighted average composition across all n sequences. Apart from PAUP*, none of these methods or programs considers the biasing effect of invariable sites, and more appropriate methods (see below) are now available.
3. Methods belonging to the third category compare the nucleotide or amino acid content in pairs of sequences using matched-pairs tests of homogeneity; therefore they are the most appropriate methods to test for violation of the assumption of stationarity. Tavaré (1986) is likely to have been the first person to have realized this, and he cited several useful papers. Among these papers, he highlighted Bowker's (1948) test for symmetry and Stuart's (1955) test for marginal symmetry, but without explaining why one might want to use both tests. Nonetheless, he did point out that it is possible to have marginal symmetry without having symmetry, thus suggesting that Bowker's (1948) test is superior to Stuart's (1955) test when violation of the assumption of stationarity is of concern. Bowker's (1948) test for symmetry has been used repeatedly (see, e.g., Lanave and Pesole, 1993; Waddell et al., 1999; Waddell and Steel, 1997), and so has Stuart's (1955)

test for marginal symmetry (see, e.g., Waddell et al., 1999); however, in the latter case, the variances and covariances were unfortunately ignored, so the test results are most likely somewhat biased. The reason why this category of methods is more appropriate than the other methods is that the matched-pairs tests consider the alignment on a site-by-site basis. A generalized version of Stuart's (1955) test for marginal symmetry is now available (Rzhetsky and Nei, 1995), but the value of this more recent version may be limited because it is still possible to have marginal symmetry without symmetry (see above); moreover, it does not allow users to identify the offending sequence(s).

4. The fourth category of methods for assessing compositional homogeneity includes a single method (Kumar and Gadagkar, 2001a, 2001b). In this method, a disparity index is calculated on the basis of (i) the frequency of the i th nucleotide (or amino acid) in sequence X , (ii) the frequency of the i th nucleotide (or amino acid) in sequence Y , and (iii) the number of sites where X and Y differ, and a Monte Carlo simulation is employed to determine whether the disparity index for the sequences is above the level of significance. The method was shown to be better than the classical χ^2 -test (i.e., one belonging to the second category of methods) but since it is not based on the two-way tests of homogeneity, it should be used with caution. Alternatively, we recommend the matched-pairs tests described above.

CONCLUDING REMARKS AND RECOMMENDATIONS

We have examined the effect of only one factor confounding phylogenetic inference, and there are many others to consider, such as rate heterogeneity among lineages and among sites, heterogeneity in the distributions and proportions of invariable sites (covarion and covarion evolution), and correlated evolution among sites. Each of these confounding factors can be regarded as separate problems in their own right but may nonetheless interact constructively or destructively in the recovery of an historical signal. They may, in fact, explain some of the discrepancies between results obtained from analyses of simulated and empirical data sets. Ho and Jermiin (2004) have so far shown that the joint effect of rate heterogeneity among lineages and compositional heterogeneity can have extremely complex, and in some cases unpredictable, effects on phylogenetic estimates, both in terms of tree topology and the estimation of edge lengths.

In the light of this, and with reference to published (e.g., Foster and Hickey, 1999; Foster et al., 1997; Jermiin et al., 1994) and unpublished research (Ho and Jermiin), we suspect the problem of compositional heterogeneity among sequences may be far more widespread than previous research has indicated; hence, we recommend that the assumption of compositional homogeneity be assessed *prior* to phylogenetic analyses of nucleotide or amino acid sequences, as already suggested by several authors (e.g., Kumar and Gadagkar, 2001b; von Haeseler et al., 1993). If the assumption of compositional

homogeneity is violated, then we recommend that phylogenetic data be analyzed using general models of nucleotide (or amino acid) substitution designed to accommodate this bias (see, e.g., Foster, 2004; Galtier and Gouy, 1995, 1998; Galtier et al., 1999; Gu and Li, 1996, 1998; Lake, 1994; Lockhart et al., 1994; Steel, 1994; Steel et al., 1993, 1995; Tamura and Kumar, 2002; Yang and Roberts, 1995).

Finally, we recommend that much stronger emphasis be given to extending the length of sequences in phylogenetic data sets, rather than increasing the number of taxa, as this will allow reliable identification of true internal edges, even when they are comparatively short.

ACKNOWLEDGMENTS

It is our pleasure to thank Ross H. Crozier, Peter G. Foster, Daniel Huson, Peter J. Lockhart, and an anonymous reviewer for constructive comments on this paper. The research was partly funded by a Discovery Grant (DP0453173) from the Australian Research Council. Simon Y. W. Ho was supported by an AE and FAQ Stephens Scholarship from the University of Sydney, Australia, whereas Faisal Ababneh was supported by a postgraduate scholarship from Al-Hussein Bin Talal University, Jordan. This is research paper #003 from SUBIT.

REFERENCES

- Baake, E., and A. von Haeseler. 1999. Distance measures in terms of substitution processes. *Theor. Popul. Biol.* 55:166–175.
- Bowker, A. H. 1948. A test for symmetry in contingency tables. *J. Am. Stat. Assoc.* 43:572–574.
- Chang, B. S. W., and D. L. Campbell. 2000. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Mol. Biol. Evol.* 17:1220–1231.
- Conant, G. C., and P. O. Lewis. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion on phylogenetic inference. *Mol. Biol. Evol.* 18:1024–1033.
- Downton, M., and A. D. Austin. 1997. The evolution of strand-specific compositional bias. A case study in the hymenopteran mitochondrial 16S rRNA gene. *Mol. Biol. Evol.* 14:109–112.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368–376.
- Felsenstein, J. 2002. PHYLIP (Phylogeny Inference Package), version 3.6(α 3). Distributed by the author.
- Fitch, W. F. 1971. Towards defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* 20:406–416.
- Foster, P. G. 2004. Modeling compositional heterogeneity. *Syst. Biol.* 53:485–495.
- Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J. Mol. Evol.* 48:284–290.
- Foster, P. G., L. S. Jermiin, and D. A. Hickey. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J. Mol. Evol.* 44:282–288.
- Galtier, N., and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl. Acad. Sci. U. S. A.* 92:11317–11321.
- Galtier, N., and M. Gouy. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogenous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15:871–879.
- Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* 283:220–221.
- Gu, X., and W.-H. Li. 1996. Bias-corrected paralogous and logdet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. *Mol. Biol. Evol.* 13:1375–1383.
- Gu, X., and W.-H. Li. 1998. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *Proc. Natl. Acad. Sci. U. S. A.* 95:5899–5905.

- Hasegawa, M., T. Hashimoto, J. Adachi, N. Iwabe, and T. Miyata. 1993. Early branching in the evolution of eukaryotes: Ancient divergence of *Entamoeba* that lacks mitochondria revealed by protein sequence data. *J. Mol. Evol.* 36:380–388.
- Hashimoto, T., Y. Nakamura, T. Kamaishi, F. Nakamura, J. Adachi, K.-I. Okamoto, and M. Hasegawa. 1995. Phylogenetic place of mitochondrial-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. *Mol. Biol. Evol.* 12:782–793.
- Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K.-I. Okamoto, and M. Hasegawa. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: Phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. *Mol. Biol. Evol.* 11:65–71.
- Ho, S. Y. W., and L. S. Jermini. 2004. Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.* 53:623–637.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
- Jermini, L. S., D. Graur, R. M. Lowe, and R. H. Crozier. 1994. Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. *J. Mol. Evol.* 39:160–173.
- Jermini, L. S., S. Y. W. Ho, F. Ababneh, J. Robinson, and A. W. D. Larkum. 2003. *Hetero*: A program to simulate the evolution of DNA on a four-taxon tree. *Appl. Bioinformatics* 2:159–163.
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 in *Mammalian protein metabolism* (H. N. Munro, ed.). Academic Press, New York.
- Klenk, H. P., P. Palm, and W. Zillig. 1994. DNA-dependent RNA polymerases as phylogenetic marker molecules. *Syst. Appl. Microbiol.* 16:638–647.
- Kumar, S., and S. R. Gadagkar. 2001a. Corrigendum—Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* 159:913–914.
- Kumar, S., and S. R. Gadagkar. 2001b. Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* 158:1321–1327.
- Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralineal distances. *Proc. Natl. Acad. Sci. U. S. A.* 91:1155–1159.
- Lanave, C., and G. Pesole. 1993. Stationary MARKOV processes in the evolution of biological macromolecules. *Binary* 5:191–195.
- Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20:86–93.
- Lanave, C., S. Tommasi, G. Preparata, and C. Saccone. 1986. Transition and transversion rate in the evolution of animal mitochondrial DNA. *Bio. Syst.* 19:273–283.
- Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. D. Larkum. 1992a. Substitutional bias confounds inference of cyanellid origins from sequence data. *J. Mol. Evol.* 34:153–162.
- Lockhart, P. J., D. Penny, M. D. Hendy, C. J. Howe, T. J. Beanland, and A. W. D. Larkum. 1992b. Controversy on chloroplast origins. *FEBS Lett.* 301:127–131.
- Lockhart, P. J., D. Penny, M. D. Hendy, and A. D. W. Larkum. 1993. Is *Prochlorothrix hollandica* the best choice as a prokaryotic model for higher plant Chl *a/b* photosynthesis? *Photosynth. Res.* 37:61–68.
- 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.
- Loomis, W. F., and D. W. Smith. 1990. Molecular phylogeny of *Dicystostelium discoideum* by protein sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 87:9093–9097.
- Olsen, G. J., and C. R. Woese. 1993. Ribosomal RNA: A key to phylogeny. *FASEB J.* 7:113–123.
- Penny, D., M. D. Hendy, E. A. Zimmer, and R. K. Hamby. 1990. Trees from sequences: Panacea or Pandora's box? *Aust. Syst. Bot.* 3:21–38.
- Preparata, G., and C. Saccone. 1987. A simple quantitative model of the molecular clock. *J. Mol. Evol.* 26:7–15.
- Rosenberg, M. S., and S. Kumar. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Mol. Biol. Evol.* 20:610–621.
- Rzhetsky, A., and M. Nei. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* 12:131–151.
- Saccone, C., G. Pezole, and G. Preparata. 1989. DNA microenvironments and the molecular clock. *J. Mol. Evol.* 29:407–411.
- Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.
- Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
- Sogin, M. L., G. Hinkle, and D. D. Leipe. 1993. Universal tree of life. *Nature* 362:795.
- Steel, M. A. 1994. Recovering a tree from the leaf colourations 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.
- Steel, M. A., P. J. Lockhart, and D. Penny. 1995. A frequency-dependent significance test for parsimony. *Mol. Phylogenet. Evol.* 4:64–71.
- Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964–969.
- Stuart, A. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42:412–416.
- Swofford, D. L. 2001. PAUP*, Phylogenetic analysis using parsimony (*and other methods). version 4. Sinauer Associates, Sunderland, Massachusetts.
- Tamura, K., and S. Kumar. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol. Biol. Evol.* 19:1727–1736.
- Tarrío, R., F. Rodriguez-Trelles, and F. J. Ayala. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. *Mol. Biol. Evol.* 18:1464–1473.
- Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.* 17:57–86.
- Van Den Bussche, R. A., R. J. Baker, J. P. Huelsenbeck, and D. M. Hillis. 1998. Base compositional bias and phylogenetic analyses: A test of the “flying DNA” hypothesis. *Mol. Phylogenet. Evol.* 13:408–416.
- von Haeseler, A., A. Janke, and S. Pääbo. 1993. Molecular phylogenetics. *Verh. Dtsch. Zool. Ges.* 86:119–129.
- Waddell, P. J., Y. Cao, 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.
- Waddell, P. J., and M. A. Steel. 1997. General time reversible distances with unequal rates across sites: Mixing Γ and inverse Gaussian distributions with invariant sites. *Mol. Phylogenet. Evol.* 8:398–414.
- Weisburg, W. G., S. J. Giovannoni, and C. R. Woese. 1989. The *Deinococcus* and *Thermus* phylum and the effect of ribosomal RNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* 11:128–134.
- Yang, Z. 1996a. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11:367–372.
- Yang, Z. 1996b. Maximum likelihood models for combining analyses of multiple sequence data. *J. Mol. Evol.* 42:587–596.
- Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branches in the tree of life. *Mol. Biol. Evol.* 12:451–458.

First submitted 1 July 2003; reviews returned 14 September 2003;

final acceptance 26 March 2004

Associate Editor: Peter Lockhart