Effects of Nucleotide Composition Bias on the Success of the Parsimony Criterion in Phylogenetic Inference

Gavin C. Conant and Paul O. Lewis†*

*Department of Biology, University of New Mexico; and †Department of Ecology and Evolutionary Biology, University of Connecticut

Convergence in nucleotide composition (CNC) in unrelated lineages is a factor potentially affecting the performance of most phylogeny reconstruction methods. Such convergence has deleterious effects because unrelated lineages show similarities due to similar nucleotide compositions and not shared histories. While some methods (such as the LogDet/paralinear distance measure) avoid this pitfall, the amount of convergence in nucleotide composition necessary to deceive other phylogenetic methods has never been quantified. We examined analytically the relationship between convergence in nucleotide composition and the consistency of parsimony as a phylogenetic estimator for four taxa. Our results show that rather extreme amounts of convergence are necessary before parsimony begins to prefer the incorrect tree. Ancillary observations are that (for unweighted Fitch parsimony) transition/transversion bias contributes to the impact of CNC and, for a given amount of CNC and fixed branch lengths, data sets exhibiting substantial site-to-site rate heterogeneity present fewer difficulties than data sets in which rates are homogeneous. We conclude by reexamining a data set originally used to illustrate the problems caused by CNC. Using simulations, we show that in this case the convergence in nucleotide composition alone is insufficient to cause any commonly used methods to fail, and accounting for other evolutionary factors (such as site-to-site rate heterogeneity) can give a correct inference without accounting for CNC.

Introduction

Since phylogenetic relationships cannot be observed, it is impossible to directly verify the accuracy of phylogeny reconstructions. Because of this difficulty, it is of interest to discover conditions in data that can be demonstrated to cause phylogeny reconstruction methods to fail. One approach has been to specify a model phylogeny and a substitution model incorporating the factor of interest and then show that data generated from that phylogeny result in incorrectly inferred relationships. This demonstration can be done analytically for simple cases and some phylogeny reconstruction methods (e.g., Felsenstein 1978), but it more often requires the use of computer simulation (e.g., Nei 1991; Kuhner and Felsenstein 1994; Huelsenbeck 1995; Schöniger and von Haeseler 1995).

For DNA sequence data, several evolutionary factors have been discovered that can potentially mislead phylogeny estimation methods. Examples of such factors include transition/transversion bias (Kimura 1980; Wakeley 1993), heterogeneity in substitution rates among lineages (Felsenstein 1978), heterogeneity in substitution rates among sites within a nucleotide sequence (Navidi, Churchill, and von Haeseler 1991; Reeves 1992; Sidow and Steel 1992; Yang 1993), nonindependence of sites within a gene (Goldman and Yang 1994; Muse 1995, 1996; Schöniger and von Haeseler 1995), and nonstationarity of nucleotide frequencies across lineages (Loomis and Smith 1990; Burggraf, Stetter, and Woese 1992; Hasegawa and Hashimoto 1993; Lockhart et al. 1994; Galtier and Gouy 1995, 1998).

Lockhart et al. (1994) presented three compelling examples in which they postulated that convergence in nucleotide composition (CNC) in independent lineages led parsimony, as well as methods based on traditional substitution models, to prefer an incorrect tree, namely the tree placing taxa with similar nucleotide compositions together. LogDet (Lake 1994; Steel 1994) was the only transformation of those tested that resulted in a correct phylogenetic inference. Relatively few other cases have been found in which CNC has been identified as a problematic factor, although Foster and Hickey (1999) suggest that it may be the cause of misleading inferences for animal phylogenies when using all mitochondrial protein-coding sequences. There are at least two plausible explanations for this paucity of examples. First, if changed nucleotide composition is inherited (fig. 1*A*) rather than acquired by convergence (fig. 1*B*), one might expect phylogeny methods such as parsimony to prefer the correct tree more strongly than they should. Thus, whether nonstationarity in nucleotide composition is a problem would depend on the relative frequency in nature of inherited versus convergent similarity in nucleotide composition. This explanation is rather difficult to investigate, as it requires ascertaining relative frequencies of inherited composition versus CNC in nature. Second, even if convergent similarity in nucleotide composition is common, whether it is a problem for phylogeny methods depends on the strength of the convergence and how CNC interacts with other evolutionary factors. In this paper, we instead concentrate on this second explanation, using analyses of four-taxon phylogenies to obtain a feeling for the amount of CNC required to mislead phylogeny methods, especially parsimony. We also present a reexamination of one of the Lockhart et al. (1994) examples using computer simulation to show that other factors are at work in these data, and CNC alone does not provide a satisfactory explanation for the failure of the phylogeny methods examined.

Key words: nucleotide composition, phylogeny, LogDet, G+C bias, maximum parsimony.

Address for correspondence and reprints: Gavin C. Conant, Department of Biology, 167 Castetter Hall, University of New Mexico, Albuquerque, New Mexico 87131-1091. E-mail: gconant@unm.edu.

Mol. Biol. Evol. 18(6):1024–1033. 2001

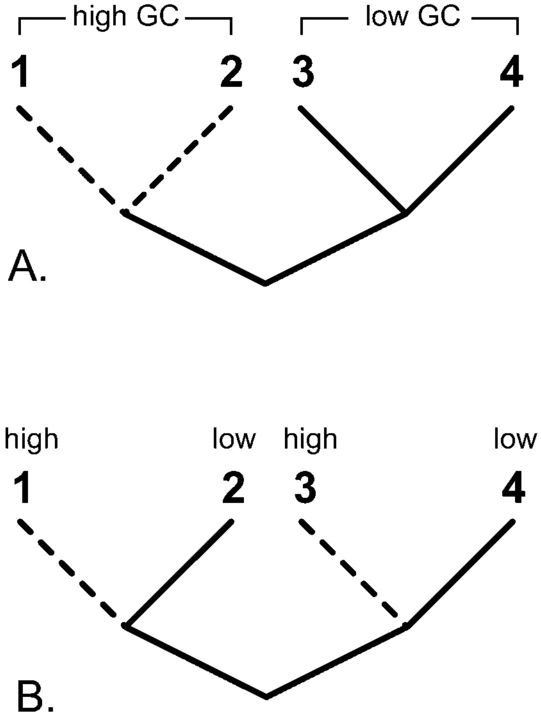

FIG. 1.—Four-taxon trees depicting different ways in which differences in G+C composition among the tip sequences can accrue. In all cases, it is assumed that an increase in the frequency with which G's and C's are recruited into sequences in the event of a substitution occurs at some point in time, and this increased propensity continues and is inherited by descendant lineages following speciation events. *A*, The increase in G and C substitutions begins in the common ancestor of sequences 1 and 2 and is inherited in these two lineages, resulting in sequences 1 and 2 having higher G+C compositions than sequences 3 and 4. *B*, The increase occurs independently in the lineage leading to sequence 1 and the lineage leading to sequence 3. For purposes of the simulations, which all used tree B as the model tree, the branch length (*d*) was identical for all branches (edges) except for the two internal segments immediately descended from the root node, for which the length was *d*/2.

Convergence in Nucleotide Composition in Four-Taxon Trees

The term ''nucleotide composition'' can have at least two distinct meanings. It can refer to the nucleotide pool available for substitution or to the observed proportions of nucleotides in a particular sequence or genome. Both have been termed ''equilibrium frequencies,'' since all commonly used substitution models (with the exception of the model underlying the LogDet/paralinear distance measure) assume that the nucleotide composition is *stationary* (i.e., does not change from lineage to lineage across the tree). We use the term ''base frequencies'' to refer to the substitution pool relative frequencies, but we allow them to change from lineage to lineage following Yang and Roberts (1995) and Galtier and Gouy (1998). When there is a change in substitution pool base frequencies, it takes some time before the observed nucleotide composition again reaches equilibrium. This lag is exacerbated by strong site-to-site rate heterogeneity, which leaves many sites unchanged for long periods of time. The appendix contains formulas for determining the expected nucleotide composition at some arbitrary time *t* following a change in base frequencies for models with and without the incorporation of rate heterogeneity.

In this section, we examine the question of how much CNC is required to mislead parsimony in the four-taxon case by using the probabilities of parsimony-informative patterns to define the region of statistical inconsistency for parsimony (i.e., the region in which parsimony would converge on an incorrect tree given an infinite amount of data). The model tree is that in figure 1*B*, consisting of two ''biased'' branches and three ''unbiased'' branches (the central branch comprises both segments attached to the root node). Because short internal branches in four-taxon trees present the greatest difficulties for phylogeny reconstruction, the length of the central branch was varied independently of the four peripheral branches. Branch lengths are given in terms of the expected number of substitutions per site (*d*) unless otherwise indicated. The K2P model (Kimura 1980) was used for unbiased branches, and the model employed for biased branches was the T92 model (Tamura 1992; Galtier and Gouy 1998). The bias introduced along the two biased branches involved increasing the frequency of both G and C by an amount δ (i.e., $\pi_G = \pi_C = 0.25 + \delta$, $\pi_A = \pi_T = 0.25 - \delta$). The probability of observing any of the four bases at the root node was assumed to be 0.25, in accordance with the K2P model employed for the central branch containing the root.

With a tree and a substitution model thus specified, it is possible to compute the probability of all 256 data patterns for any given combination of G+C bias (δ), transition/transversion rate ratio (κ), and branch length (*d*). We need be concerned with only 36 of the 256 possible patterns, 12 of which support each of the three possible unrooted trees. Let P_0 be the sum of the probabilities of the 12 patterns supporting the true tree and let P_1 and P_2 be the sum of the probabilities of the 12 patterns supporting each of the two incorrect trees. If either P_1 or P_2 exceeds P_0, then parsimony will tend to choose incorrectly even with an infinite number of nucleotide sites (i.e., parsimony is statistically inconsistent).

As expected, for many combinations of branch lengths and κ, increasing G+C bias (δ) caused parsimony to become statistically inconsistent (fig. 2). Since the model tree specified the biased branches to be those leading to sequences 1 and 3, the tree that placed sequences 1 and 3 (tree 1) together was increasingly supported as the level of bias increased. Tree 0 (the true tree, placing sequences 1 and 2 together) and tree 1 thus provided the comparison of interest; tree 2 (placing sequences 1 and 4 together) will be ignored hereinafter. The plots in figure 2 depict the difference between P_0 and P_1. The region of inconsistency (shaded) is entered when the surface representing $P_0 - P_1$ dips below 0; it is in this area that parsimony is expected to prefer tree 1 over the true tree.

It has been suggested by Lockhart et al. (1992) that statistical inconsistency as a result of CNC occurs in the four-taxon case only when the internal branch of the unrooted tree is shorter than the terminal branches. This

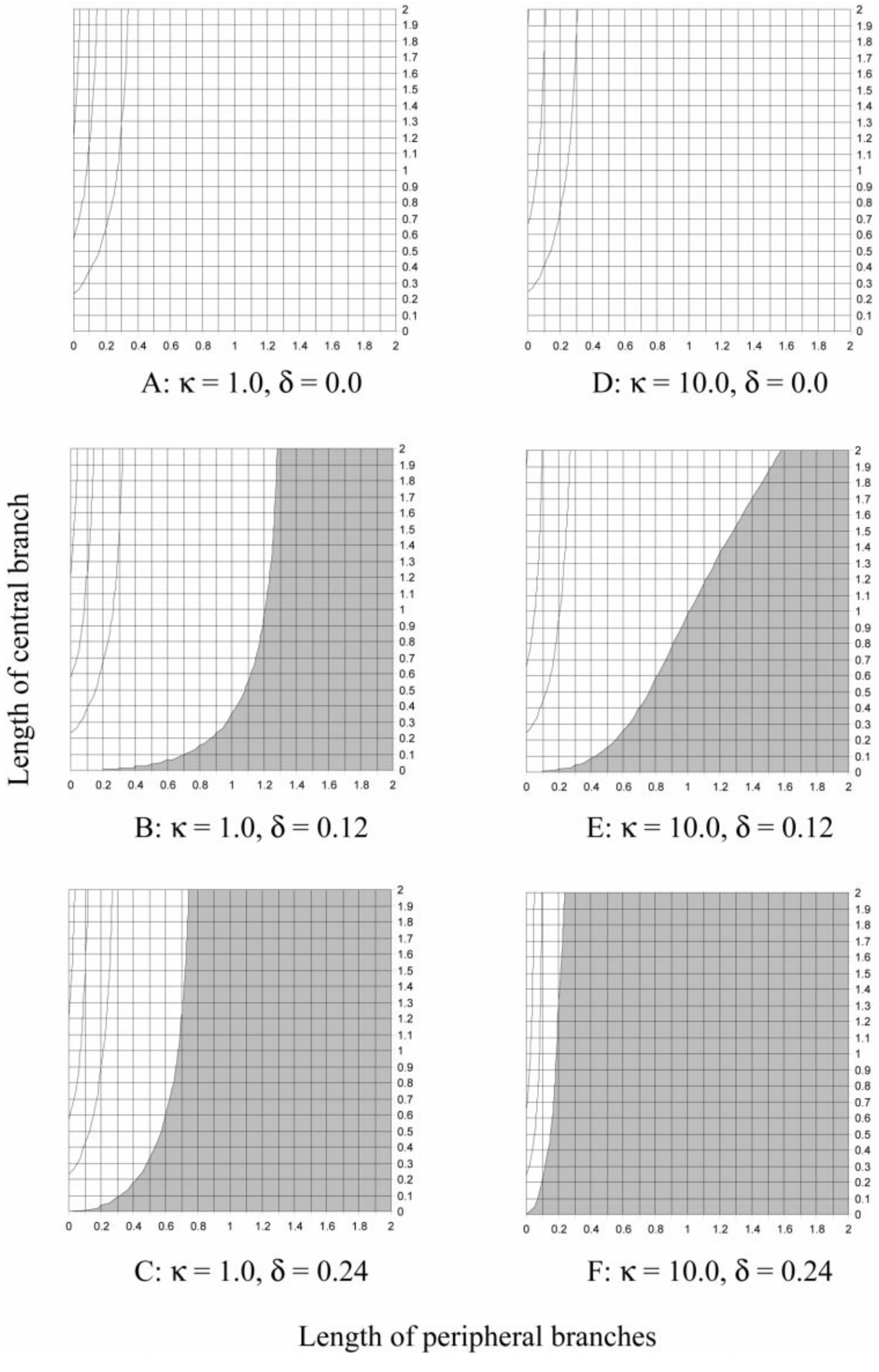

FIG. 2.—Expected performance of the parsimony criterion for differing combinations of *d*, κ, and δ, where *d* represents the expected number of substitutions per site, κ is the transition/transversion rate ratio (the instantaneous transition rate divided by the instantaneous transversion rate), and δ is the magnitude of the increase in the equilibrium frequencies of both G and C ($\pi_G = \pi_C = 0.25 + \delta$, $\pi_A = \pi_T = 0.25 - \delta$) on biased branches (the dashed lines in the tree depicted in fig. 1*B*). The performance of parsimony is measured as the difference between the probability of observing data patterns that support the correct tree and the probability of data patterns that support the ''G+C tree'' (i.e., the tree that incorrectly places taxa with increased G+C content together). Shaded portions of the plots represent regions of statistical inconsistency for parsimony, analogous to the ''Felsenstein Zone'' in the long-branch attraction problem, since in these regions misleading data patterns are more probable than patterns supporting the correct tree. *A*, κ equals 1.0, δ = 0.0. *B*, κ equals 1.0, δ = 0.12. *C*, κ equals 1.0, δ = 0.24. *D*, κ equals 10.0, δ = 0.0. *E*, κ equals 10.0, δ = 0.12. *F*, κ equals 10.0, δ = 0.24.

suggests that CNC is a problem related to the long-branch attraction described by Felsenstein (1978). The vertical axis in figure 2 represents the length of the central branch, while the horizontal axis represents the length of each of the peripheral branches. Figure 2 demonstrates that, in fact, even if the internal branch is equal in length to the terminal branches, there exists a level of G+C bias sufficient to cause parsimony to become inconsistent, although the level of bias required in such cases is quite high.

Figure 2 shows that, in general, branch lengths must be large (>0.5 substitutions per site) for CNC to cause serious problems for parsimony, even when the G+C bias is nearly at its maximum possible value ($\delta = 0.24$). CNC is exacerbated by small internal branch lengths and especially by transition/transversion bias.

Figure 3 repeats the analysis of figure 2, this time including the discrete gamma distribution of sitewise relative rates. In this case, we see that the addition of rate heterogeneity actually decreases the size of the zone of inconsistency, especially in regions where all branches are long. One might predict that site-to-site rate heterogeneity would make matters worse for parsimony (and any method that does not take it into account), since high rate heterogeneity implies that change is concentrated at fewer sites. This means that variable sites have a better chance of experiencing multiple hits than in the rate homogeneity case, leading to greater difficulty in distinguishing true phylogenetic signal from false signal due to convergence. This would be especially true if the total amount of accumulated nucleotide composition bias were held constant. In figure 2, this is not the case: it is the number of substitutions (branch lengths) that is held constant, and the greater success of parsimony can thus be attributed to the fact that change has been concentrated at a few variable sites, and the realized nucleotide composition bias is not as great as that for the rate homogeneity case (where more sites have undergone at least one change).

Simulation Study

The rigidity of the model tree in the analytical study makes it difficult to apply the results to real data sets. In particular, few real data sets follow the assumed perfect molecular clock, and fewer still have interior nodes so evenly spaced in time. We therefore used computer simulation to study the effects of CNC on the ability of parsimony and other methods to reconstruct the true tree using the chlorop.phy data set obtained from http://imbs.massey.ac.nz/Research/MolEvol/Farside/programs.htm and described in Lockhart et al. (1994). Lockhart et al. (1994) examined data from the 16S rRNA gene of chloroplasts (of diverse phylogenetic origins), as well as the cyanobacterium Anacystis. They showed that many common phylogenetic reconstruction methods failed to favor the tree assumed to be correct, which places all the chlorophyll b–containing organisms together, separated from the cyanobacterium Anacystis and the chlorophyll c–containing chromophyte alga Olithodiscus. The methods that failed were (1) parsimony, presumably equal-weighted and using unordered character states; (2) maximum likelihood, using the model described in Felsenstein (1993), presumably with the transition : transversion ratio fixed at the default value of 2; (3) neighbor joining using Jukes and Cantor (1969) distances; and (4) neighbor joining using Kimura (1980) two-parameter distances. These methods all placed Euglena between Anacystis and Olithodiscus. Using the LogDet transformation (in conjunction with neighbor joining) on just parsimony-informative sites produced the well-corroborated tree in which Euglena grouped with the other chlorophyll a/b–containing organisms. Lockhart et al. (1994) concluded that the relatively low G+C content of Euglena and Olithodiscus caused most methods to group them together.

Using PAUP*, version 4.0d64 (Swofford 1998), we were able to reproduce the results of Lockhart et al. (1994) on the entire data matrix of eight sequences, but we reduced the data set to just the sequences from Anacystis, Olithodiscus, Euglena, and Chlamydomonas for simplicity. As table 1 shows, reducing the taxon sampling did not affect the general conclusions reached by Lockhart et al. (1994). All methods examined except LogDet favored the unrooted tree topology grouping Euglena and Olithodiscus and separating them from Chlorella and Anacystis, which have higher G+C contents (table 2). The model described by Galtier and Gouy (1998), hereinafter called the GG98 model, was used to simulate data according to the tree presumed to be correct, namely, (Anacystis, Olithodiscus, (Euglena, Chlamydomonas)). In essence, the hypothesis tested was that the process underlying the evolution of the observed sequences did not differ from the model of evolution used in the simulations. The results of the previous section suggest that the degree of bias present in the Lockhart et al. (1994) data set is not large enough to mislead parsimony (or, presumably, other methods) unless other factors exacerbate its effects. We therefore predicted that all methods would usually pick the correct tree in the simulated data sets.

The parameter values used in the simulations were maximum-likelihood estimates obtained using two independently written computer programs, each using the GG98 model. The program EVAL_NH, written by Galtier and Gouy, was used to check the results from a program (GG98) written separately by one of us (P.O.L.). It is important to note that the incorporation of CNC makes the model non-time-reversible. In such models, the maximum likelihood changes with different rootings, so table 3 presents likelihood scores for all 15 possible rooted topologies for four taxa. The maximum-likelihood tree under the GG98 model is the ‘‘true’’ tree (table 3). This result demonstrates that using a model allowing nucleotide composition to vary across the tree improves the quality of the estimated tree. The two programs were in agreement with respect to the parameter estimates for the maximum-likelihood tree (fig. 4). We each wrote independent computer programs to simulate data sets based on these parameter estimates and used PAUP*, version 4.0d64 (Swofford 1998), to evaluate each of the 1,000 simulated data sets for the five methods used by Lockhart et al. (1994) and described above:

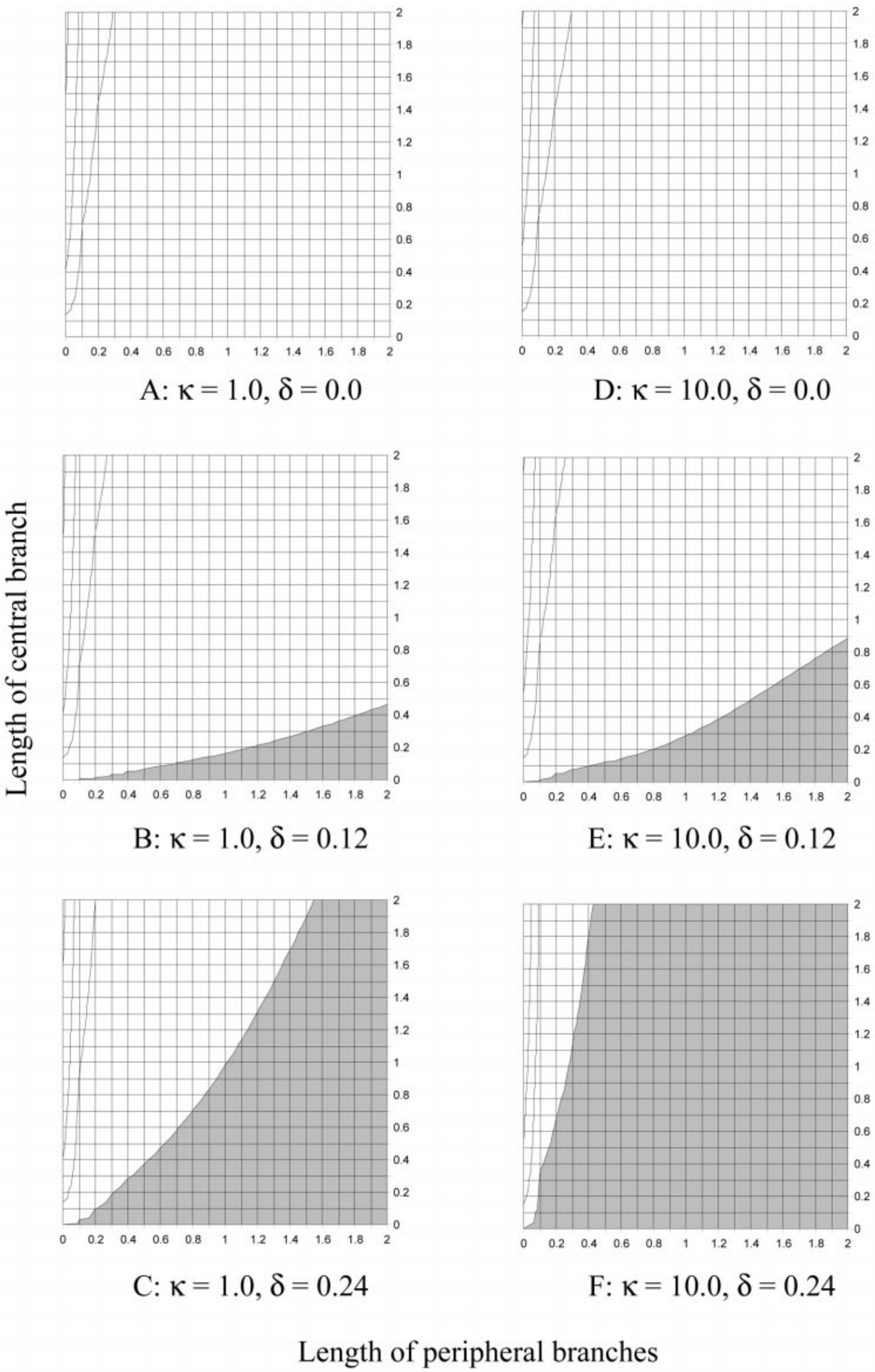

FIG. 3.—Plot of the performance of parsimony as in figure 2, with the addition of site-to-site rate variation modeled as a discrete gamma distribution with four categories and α (gamma shape parameter) = 0.2. *A,* κ equals 1.0, δ = 0.0. *B,* κ equals 1.0, δ = 0.12. *C,* κ equals 1.0, δ = 0.24. *D,* κ equals 10.0, δ = 0.0. *E,* κ equals 10.0, δ = 0.12. *F,* κ equals 10.0, δ = 0.24.

Table 1
Performance of Various Phylogenetic Inference Methods on the Eight-Taxon Data Set of Lockhart et al. (1994) and with the Four Taxa Subsequently Used

	Eight-Taxon Data Set		Four-Taxon Data Set	
Method	Score	Correct?	Score	Correct?
MP	393	No	280	No
ML-F84[a]	−3,158.20339	No	−2,485.78542	No
ML-F84 (tratio est.)	−3,158.19033	No	−2,485.74291	No
ML-F84 (shape est.)	−3,067.68322	Yes	−2,458.67618	Yes
ML-F84 (tratio and shape est.)	−3,065.27760	Yes	−2,457.56807	Yes
ME-JC	0.42234	No	0.31589	No
ME-JC (shape = MLE)[b]	0.60244	No	0.42556	No
ME-K2P	0.42693	No	0.32001	No
ME-K2P (shape = MLE)	0.66477	No	0.48482	No
ME-LogDet	0.43550	Yes	0.32255	Yes

Note.—The liverwort, tobacco, rice, and Chlorella sequences were removed to create the four-taxon data set. MP = maximum parsimony; ML = maximum likelihood; JC = Jukes-Cantor distances; K2P = Kimura two-parameter distances.

[a] Transition/transversion ratio = 2.0 in ML analyses unless ''tratio est.'' is specified; rate homogeneity is assumed unless ''shape est.'' is specified.

[b] Maximum-likelihood estimate of gamma shape parameters using the same substitution model and assumed discrete gamma distribution with four rate categories.

equal-weighted parsimony (MP); maximum likelihood with the F84 model (ML); minimum evolution with Jukes and Cantor (1969) distances (ME-JC); minimum evolution with K2P distances (ME-K2P); and minimum evolution with LogDet distances (ME-LogDet). None of these methods selected an incorrect tree in any of the 1,000 simulations, suggesting that there is a significant difference between the model used for simulation and the actual processes generating the observed sequences.

We repeated the simulations, this time incorporating discrete gamma rate heterogeneity into the data. The model used is termed the GG98-Γ model, as it is identical to the GG98 model except for the addition of a gamma shape parameter. Four rate categories were used, with the mean of each category serving as the relative rate used in the likelihood calculations. Again, when the likelihood of each of the 15 possible rooted trees was computed using the GG98-Γ model, the maximum-likelihood tree was identical to the tree topology assumed to be true by Lockhart et al. (1994) (table 3). The maximum-likelihood estimates of the parameters of the GG98-Γ model (fig. 5) were used as the basis of the simulations; however, this time only the GG98 program could be used to estimate parameters because EVAL_NH does not include the gamma version of the GG98 model. In this case, some of the simulated data sets resulted in incorrect estimates of phylogeny regardless of the method used. Nevertheless, all of the methods recovered the correct tree a high percentage of the time, and LogDet did not outperform the other methods (table 4) when presented with the true amount of rate heterogeneity (the maximum-likelihood estimate of the gamma shape parameter from the original data set, 0.308, was the assumed level of rate heterogeneity in the simulated data).

Discussion

Of the many evolutionary factors affecting the accuracy of phylogenetic inference, CNC is a relative newcomer, being recognized formally as a problem with the papers by Lake (1994), Lockhart et al. (1992), and Steel (1994). The present paper seeks to discover how much CNC is required before it presents serious problems for phylogenetic inference methods such as parsimony. The analytical results presented suggest that extreme combinations of substitution rates, transition/transversion bias, and equilibrium frequencies are required before parsimony is expected to fail. This is welcome news, because the situation investigated here represents nearly the worst-case scenario: nucleotide composition con-

Table 2
Results of Analysis of Four Taxa from Lockhart et al. (1994) Under Different Inference Methods

Tree	ML[a]	MP[b]	NJ-JC[c]	NJ-K2P[d]	NJ-LogDet[e]
(A, B, (C, D))	−2,486.40247	282	0.31591	0.32010	0.32255
(A, C, (B, D))	−2,500.71061	294	0.35009	0.35590	0.35478
(A, D, (B, C))	−2,485.78542	280	0.31589	0.32001	0.32373

Note.—Taxon abbreviations: *A* = Anacystis; B = Olithodiscus; C = Euglena; D = Chlamydomonas. All computations were performed with PAUP*, version 4.0d64 (Swofford 1998).

[a] Natural logarithm of maximum likelihood (F84 model, empirical base frequencies and transition/transversion ratio = 2.0).

[b] Maximum-parsimony tree length (unordered characters and equally weighted character state transitions).

[c] Neighbor joining using Jukes-Cantor distances, minimum evolution score.

[d] Neighbor joining using Kimura two-parameter distances, minimum evolution score.

[e] Neighbor joining using LogDet distances, minimum evolution score.

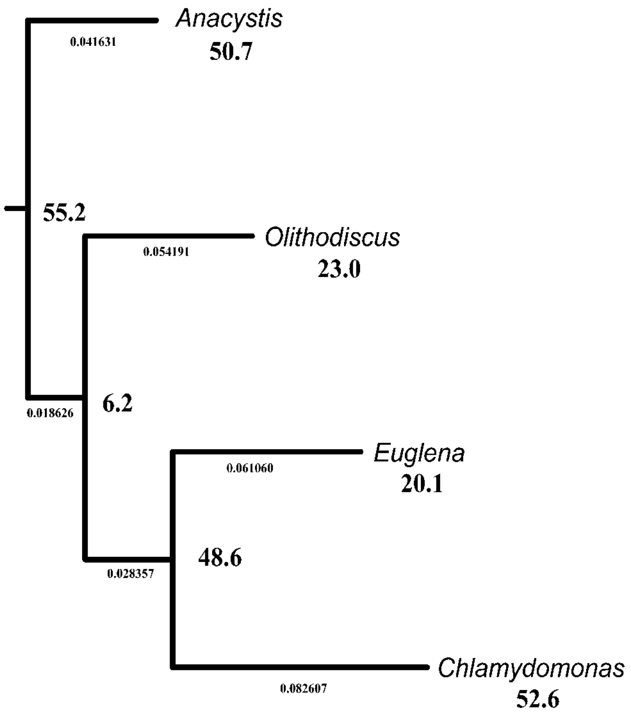

FIG. 4.—Parameter values used for the simulations using the GG98 model. These values represent maximum-likelihood estimates obtained using the GG98 model. Values below branches represent branch lengths computed using the standard HKY85 model formula for expected number of substitutions. Note that since the ''equilibrium frequencies'' differ for each branch in the GG98 model, the standard formula no longer reflects the expected number of substitutions, since the nucleotide composition is nonstationary. The correct formulas for this case are presented in the appendix. Numbers to the right of each node (or below taxon names) are the estimated percentages of G+C for the branch subtending the node. These do not represent the G+C composition of the sequence at the node, but instead represent the probabilities of substitution of G's and C's over the life span of the lineage leading up to the node. The estimated value of κ in this case was 3.781608.

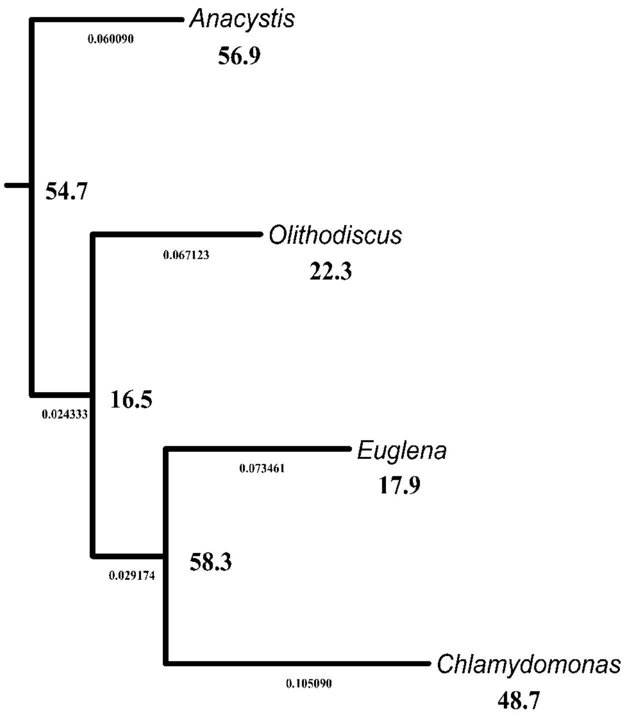

FIG. 5.—Maximum-likelihood estimates of parameters obtained under the GG98-Γ model. The values below branches and beside nodes have the same meanings as in figure 4. The estimates for κ and the gamma shape parameter are 4.673279 and 0.307850, respectively.

verging toward a common value in two unrelated lineages (the worst-case scenario for the four-taxon problem would involve increases in G+C in two unrelated terminal lineages and a corresponding decrease in G+C in the other two terminal lineages). Inherited similarities in nucleotide composition, on the other hand, will not be as problematic, as parsimony will tend to estimate trees correctly, albeit for the wrong reason. The only drawback posed by inherited similarities in nucleotide composition will be a tendency for parsimony to prefer the correct tree more strongly than it should, exhibiting a false degree of confidence in the form of bootstrap or decay values (Swofford et al. 2001).

Few clear cases have been reported in which CNC has been thought to derail the phylogenetic inference process. Of the three cases presented by Lockhart et al. (1994), two involve 18S rDNA from vertebrates and COII mtDNA from honeybees. In these two data sets, we could not find any way to obtain the putative ''correct'' tree except by using LogDet/paralinear distances, as reported by Lockhart et al. (1994). It is notable, however, that it is necessary to exclude all constant and autapomorphic sites (analyzing only parsimony-informative sites) to accurately estimate the phylogeny for these data sets. This suggests site-to-site rate heterogeneity as the likely culprit; however, taking account of site-to-site rate heterogeneity using the standard methods fails to produce a correct estimate. Therefore other, as yet unidentified, factors must be at work in these data sets.

Table 3
Natural Log of the Likelihood for the 15 Possible Rooted Trees from Lockhart et al. (1994) for the Galtier and Gouy (1998) Model With and Without Discrete Gamma Rate Heterogeneity

Tree	GG98[a]	GG98-Γ[b]
((A, B), (C, D))	−2,460.608305	−2,430.073634
((A, D), (C, B))	−2.460.299516	−2,431.704680
((A, C), (D, B))	−2,474.811480	−2,434.132779
(A, (B, (C, D)))[c]	−2,457.770740	−2,428.829561
(B, (A, (C, D)))	−2,460.911565	−2,430.073634
(C, (D, (A, B)))	−2,460.000532	−2,428.909027
(D, (C, (A, B)))	−2,459.423790	−2,429.815718
(A, (C, (B, D)))	−2,472.723745	−2,433.514116
(C, (A, (B, D)))	−2,473.956835	−2,432.968496
(B, (D, (A, C)))	−2,474.946236	−2,434.089933
(D, (B, (A, C)))	−2,473.359449	−2,434.039427
(A, (D, (B, C)))	−2,459.887635	−2,431.355763
(D, (A, (B, C)))	−2,460.380030	−2,431.676526
(B, (C, (A, D)))	−2,464.581165	−2,432.304177
(C, (B, (A, D)))	−2,463.186408	−2,430.966041

NOTE.—Taxon abbreviations: A = Anacystis; B = Olithodiscus; C = Euglena; D = Chlamydomonas.

[a] Galtier and Gouy (1998) model.

[b] GG98 model with discrete gamma rate heterogeneity (four rate categories).

[c] Maximum-likelihood topology for both models.

Table 4
Results of Simulations Based on Parameter Estimates Made Using the GG98-Γ Model

Optimality Criterion	Model	% Correct[a]
Maximum parsimony	Equal weights	94.2
Maximum likelihood	F84	96.2
Minimum evolution	JC69	92.4
	K80	91.5
	LogDet	89.8

[a] Out of 1,000 simulation replicates.

The simulation study reported here represents a test of the hypothesis that CNC alone, or CNC in combination with site-to-site rate heterogeneity, is sufficient to explain the failure of many phylogenetic methods for the third case presented by Lockhart et al. (1994) (represented by the chlorop.phy data set). We used a parametric bootstrap approach in which parameters were estimated from the data using maximum likelihood and simulations performed using these parameter estimates. The results show that CNC, either alone or in combination with site-to-site rate heterogeneity, is insufficient to account for difficulties found in the original data set. None of the simulated data sets presented problems for parsimony or any of the other methods tested (all of which failed on the original data set).

It is clear that the GG98 model used for the simulations did not capture some factor important in the evolution of the actual sequences. One possibility is that the GG98 model does not allow enough variation in nucleotide composition across the tree. This model places some constraints on changes in nucleotide composition, forcing the frequency of G to equal the frequency of C and allowing only changes in G+C composition at the nodes of the tree. It seems unlikely that these two model constraints can account for the differences seen between the simulation results and the results from the original data. First, allowing the composition of G to differ from the composition of C should not increase the chances of an artifactual joining of Euglena to Olithodiscus, since it is the low G+C content in these lineages that is postulated to have caused problems in the original data set. Second, allowing nucleotide composition to vary within lineages should also not increase the chance of Euglena pairing with Olithodiscus, since all of the phylogenetic methods that failed on the original data set view branches as the smallest units making up a phylogenetic tree: that is, they cannot, like LogDet, take account of changes in composition that occur within branches.

When simulations incorporated both CNC and rate heterogeneity, a small fraction of the simulated data sets proved difficult for all methods. This falls short of the result that would be expected if rate heterogeneity were the all-important missing factor. Also, we would expect LogDet to perform well (as it did on the original data set) compared with the other methods examined. In fact, LogDet behaves similarly to the other methods, failing on a small fraction of the simulated data sets (table 4). These observations indicate the presence of as-yet-unknown evolutionary factors at work in the evolution of the actual sequences that are not being modeled by the simulations.

The phylogenetic methods in common use today each have their own ‘‘Achilles’ heel,’’ and it behooves researchers to learn as much as possible about the factors at work in their data prior to deciding on a method to use in the final analysis. For example, parsimony’s primary Achilles’ heel has long been identified as long-branch attraction (Felsenstein 1978). Maximum likelihood can correct for problems that are identified and incorporated into substitution models but can be deceived by factors not represented in the model used (e.g., rate heterogeneity; Gaut and Lewis 1995). This paper has addressed a potential Achilles’ heel applicable to most methods of phylogenetic inference and found that it is perhaps not as great a threat as it was initially perceived to be. This is not to say that CNC can be ignored altogether. Figure 3 illustrates that CNC in combination with site-to-site rate heterogeneity and transition/transversion bias can cause problems even at biologically realistic substitution rates and levels of rate heterogeneity. For example, in figure 3, one point at which parsimony is inconsistent is characterized by the following parameter values: peripheral branch lengths = 0.8, central branch length = 0.1, gamma shape = 0.2, and transition/transversion rate ratio = 1.0, with a G+C difference of 0.12 between biased and unbiased lineages. These branch lengths and the G+C bias are at the edge of what is normally observed in actual data sets, but none are out of the realm of possibility, and the transition/transversion bias and degree of rate heterogeneity are not at all extreme. LogDet/paralinear distances provide a practical means for diagnosing CNC should it be present in a dosage sufficient to cause problems. A tree estimated using LogDet that differs from trees estimated using other methods should prompt an examination of the data for evidence that other methods are incorrectly joining taxa with similar nucleotide compositions.

While it is unlikely that any data set can be found that shows the influence of one and only one evolutionary factor, it is nevertheless beneficial to thoroughly analyze sequence data sets in the search for good examples of the effects of evolutionary factors representing potential pitfalls for phylogeny methods. Equally important is the search for new evolutionary factors. It is only when such evolutionary factors as site-to-site rate heterogeneity, transition/transversion bias, evolutionary dependence among sites, and CNC are discovered that work can begin on creating evolutionary models that avoid the problems they create.

Acknowledgments

The authors would like to thank Peter Lockhart, Michael Steel, Michael Hendy, and David Penny for making the data sets used in their 1994 paper freely available to other researchers over the World Wide Web. Permission from David L. Swofford to use a prerelease test version of his software PAUP*, version 4.0, is also gratefully acknowledged. Finally, we thank the Biology

Department of the University of New Mexico for providing support for the computing facilities needed to carry out this research. P.O.L. gratefully acknowledges funding from the Alfred P. Sloan Foundation/National Science Foundation (grant 98-4-5 ME). This paper is the culmination of the research performed for a Senior Honors Thesis by G.C.C.

APPENDIX

The transition probabilities for the HKY85 model are

$$P_{ij}(t) = \begin{cases} \pi_j + \pi_j\left(\dfrac{1}{\Pi_j} - 1\right)e^{-\beta t} + \left(\dfrac{\Pi_j - \pi_j}{\Pi_j}\right)e^{-\beta t[1+\Pi_j(\kappa-1)]} \\ \qquad\qquad i = j \\ \pi_j + \pi_j\left(\dfrac{1}{\Pi_j} - 1\right)e^{-\beta t} - \left(\dfrac{\pi_j}{\Pi_j}\right)e^{-\beta t[1+\Pi_j(\kappa-1)]} \\ \qquad\qquad i \neq j, \quad \text{transition} \\ \pi_j(1 - e^{-\beta t}) \qquad i \neq j, \quad \text{transversion,} \end{cases}$$

where π_i is the substitution pool frequency of base *i*, Π_i is the substitution pool frequency of base *i*'s group (i.e., the frequency of either purines or pyrimidines), β is the instantaneous substitution rate, κ is the transition/transversion rate ratio, and t is time. The nucleotide composition of base j after time t may be found as follows:

$$\pi_j(t) = \sum_{i\in\{\mathrm{A,C,G,T}\}} \pi_i^{(0)}P_{ij}(t).$$

Thus, the composition of A after time t is

$$\begin{aligned} \pi_{\mathrm{A}}(t) &= \pi_{\mathrm{A}}^{(0)}\left\{\pi_{\mathrm{A}} + \pi_{\mathrm{A}}\left(\frac{1}{\pi_{\mathrm{R}}} - 1\right)e^{-\beta t} + \left(\frac{\pi_{\mathrm{A}}}{\pi_{\mathrm{R}}}\right)e^{-\beta t[1+\pi_{\mathrm{R}}(\kappa-1)]}\right\} \\ &\quad + \pi_{\mathrm{G}}^{(0)}\left\{\pi_{\mathrm{A}} + \pi_{\mathrm{A}}\left(\frac{1}{\pi_{\mathrm{R}}} - 1\right)e^{-\beta t} - \left(\frac{\pi_{\mathrm{A}}}{\pi_{\mathrm{R}}}\right)e^{-\beta t[1+\pi_{\mathrm{R}}(\kappa-1)]}\right\} \\ &\quad + \pi_{\mathrm{C}}^{(0)}\pi_{\mathrm{A}}(1 - e^{-\beta t}) + \pi_{\mathrm{T}}^{(0)}\pi_{\mathrm{A}}(1 - e^{-\beta t}) \\ &= \pi_{\mathrm{A}} + \pi_{\mathrm{A}}\left(\frac{\pi_{\mathrm{R}} - \pi_{\mathrm{R}}^{(0)}}{\pi_{\mathrm{R}}}\right)e^{-\beta t} \\ &\quad + \left(\frac{\pi_{\mathrm{A}}^{(0)}\pi_{\mathrm{R}} - \pi_{\mathrm{R}}^{(0)}\pi_{\mathrm{A}}}{\pi_{\mathrm{R}}}\right)e^{-\beta t[1+\pi_{\mathrm{R}}(\kappa-1)]}, \end{aligned}$$

which reduces to the formula corresponding to the F81 model, $\pi_{\mathrm{A}} + (\pi_{\mathrm{A}}^{(0)} - \pi_{\mathrm{A}})e^{-\beta t}$ when there is no transition bias ($\kappa = 1$). More generally,

$$\pi_i(t) = \pi_i + \pi_i\left(\frac{\Pi_i - \Pi_i^{(0)}}{\Pi_i}\right)e^{-\beta t} + \left(\frac{\pi_i^{(0)}\Pi_i - \Pi_i^{(0)}\pi_i}{\Pi_i}\right)e^{-\beta t[1+\Pi_i(\kappa-1)]}.$$

Following Waddell and Steel (1997), the expected value of $\pi_i(t)$ when rates vary over sites according to a gamma distribution can be found by substituting $[1 + (\beta t/\alpha)]^{-\alpha}$ for $e^{-\beta t}$ and $\{1 + [\beta t(1 + \Pi_i(\kappa - 1))/\alpha]\}^{-\alpha}$ for $e^{-\beta t[1+\Pi_i(\kappa-1)]}$:

$$\pi_i(t) = \pi_i + \pi_i\left(\frac{\Pi_i - \Pi_i^{(0)}}{\Pi_i}\right)\left(1 + \frac{\beta t}{\alpha}\right)^{-\alpha} + \left(\frac{\pi_i^{(0)}\Pi_i - \Pi_i^{(0)}\pi_i}{\Pi_i}\right)\left(1 + \frac{\beta t[1 + \Pi_i(\kappa - 1)]}{\alpha}\right)^{-\alpha}.$$

The transition probabilities, as well as $\pi_i(t)$, for the F84 model can be obtained from those of the HKY85 model by substituting $\kappa + 1$ for the quantity $1 + \Pi_j(\kappa - 1)$.

LITERATURE CITED

Burggraf, S. G., K. O. Stetter, and C. R. Woese. 1992. A phylogenetic analysis of *Aquifex pyrophilus.* Syst. Appl. Microbiol. **15**:352–356.

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

———. 1993. PHYLIP (phylogeny inference package). Version 3.5. Distributed by the author, Department of Genetics, University of Washington, Seattle, Washington.

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.

Galtier, N., and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA **92**:11317–11321.

———. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15**:871–879.

Gaut, B., and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. **12**:152–162.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

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

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

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

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

Kuhner, M. K., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

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

LOCKHART, P. J., D. PENNY, M. D. HENDY, C. J. HOWE, T. J. BEANLAND, and A. W. D. LARKUM. 1992. Controversy on chloroplast origins. FEBS Lett. **301**:127–131.

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 Dictyostelium discoideum by protein sequence comparison. Proc. Natl. Acad. Sci. USA **87**:9093–9097.

MUSE, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139**: 1429–1439.

———. 1996. Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. **13**:105–114.

NAVIDI, W. C., G. A. CHURCHILL, and A. VON HAESELER. 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. **8**:128–143.

NEI, M. 1991. Relative efficiencies of different treemaking methods for molecular data. Pp. 90–128 *in* M. M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

REEVES, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. **35**:17–31.

SCHÖNIGER, M., and A. VON HAESELER. 1995. Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence sites are not independent. Syst. Biol. **44**:533–547.

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

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

SWOFFORD, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0 (prerelease test version). Sinauer, Sunderland, Mass.

SWOFFORD, D. L., P. J. WADDELL, J. P. HUELSENBECK, P. G. FOSTER, P. O. LEWIS, and J. S. ROGERS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. (in press).

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. **9**:678–687.

WADDELL, P., and M. STEEL. 1997. General time-reversible distances with unequal rates across sites: mixing G and inverse Gaussian distributions with invariant sites. Mol. Phylogenet. Evol. **8**:398–414.

WAKELEY, J. 1993. Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:426–442.

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., and D. ROBERTS. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. **12**:451–458.

MASAMI HASEGAWA, reviewing editor

Accepted February 13, 2001