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), non-independence 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. 1A) rather than acquired by convergence (fig. 1B), 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-taxa 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.
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 1B, 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., πC = πc = 0.25 + δ, πA = πT = 0.25 − δ). 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 P0 be the sum of the probabilities of the 12 patterns supporting the true tree and let P1 and P2 be the sum of the probabilities of the 12 patterns supporting each of the two incorrect trees. If either P1 or P2 exceeds P0, 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 P0 and P1. The region of inconsistency (shaded) is entered when the surface representing P0 − P1 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 $d$, $\kappa$, and $\delta$, where $d$ represents the expected number of substitutions per site, $\kappa$ is the transition/transversion rate ratio (the instantaneous transition rate divided by the instantaneous transversion rate), and $\delta$ 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. 1B). 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, $\kappa$ equals 1.0, $\delta = 0.0$. B, $\kappa$ equals 1.0, $\delta = 0.12$. C, $\kappa$ equals 1.0, $\delta = 0.24$. D, $\kappa$ equals 10.0, $\delta = 0.0$. E, $\kappa$ equals 10.0, $\delta = 0.12$. F, $\kappa$ equals 10.0, $\delta = 0.24$. 

Length of central branch

A: $\kappa = 1.0, \delta = 0.0$

B: $\kappa = 1.0, \delta = 0.12$

C: $\kappa = 1.0, \delta = 0.24$

D: $\kappa = 10.0, \delta = 0.0$

E: $\kappa = 10.0, \delta = 0.12$

F: $\kappa = 10.0, \delta = 0.24$

Length of peripheral branches
suggests that CNC is a problem related to the longbranch 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 chlorophy 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 due to the relatively low $G+C$ content of Euglena and Olithodiscus, 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 a (gamma shape parameter) = 0.2. A, $\kappa$ equals 1.0, $\delta = 0.0$. B, $\kappa$ equals 1.0, $\delta = 0.12$. C, $\kappa$ equals 1.0, $\delta = 0.24$. D, $\kappa$ equals 10.0, $\delta = 0.0$. E, $\kappa$ equals 10.0, $\delta = 0.12$. F, $\kappa$ equals 10.0, $\delta = 0.24$. 

Length of central branch

Length of peripheral branches
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).

**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

<table>
<thead>
<tr>
<th>Method</th>
<th>Eight-Taxon Data Set</th>
<th>Score</th>
<th>Correct?</th>
<th>Four-Taxon Data Set</th>
<th>Score</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td></td>
<td>393</td>
<td>No</td>
<td></td>
<td>280</td>
<td>No</td>
</tr>
<tr>
<td>ML-F84</td>
<td></td>
<td>−3,158.20339</td>
<td>No</td>
<td></td>
<td>−2,458.75842</td>
<td>No</td>
</tr>
<tr>
<td>ML-F84 (tratio est.)</td>
<td></td>
<td>−3,158.19033</td>
<td>No</td>
<td></td>
<td>−2,485.74291</td>
<td>No</td>
</tr>
<tr>
<td>ML-F84 (shape est.)</td>
<td></td>
<td>−3,067.68322</td>
<td>Yes</td>
<td></td>
<td>−2,458.67618</td>
<td>Yes</td>
</tr>
<tr>
<td>ML-F84 (tratio and shape est.)</td>
<td></td>
<td>−3,065.27760</td>
<td>Yes</td>
<td></td>
<td>−2,457.56807</td>
<td>Yes</td>
</tr>
<tr>
<td>ME-JC</td>
<td></td>
<td>0.42234</td>
<td>No</td>
<td></td>
<td>0.31589</td>
<td>No</td>
</tr>
<tr>
<td>ME-JC (shape = MLE)b</td>
<td></td>
<td>0.60244</td>
<td>No</td>
<td></td>
<td>0.42556</td>
<td>No</td>
</tr>
<tr>
<td>ME-K2P</td>
<td></td>
<td>0.42693</td>
<td>No</td>
<td></td>
<td>0.32001</td>
<td>No</td>
</tr>
<tr>
<td>ME-K2P (shape = MLE)</td>
<td></td>
<td>0.66477</td>
<td>No</td>
<td></td>
<td>0.48482</td>
<td>No</td>
</tr>
<tr>
<td>ME-LogDet</td>
<td></td>
<td>0.43550</td>
<td>Yes</td>
<td></td>
<td>0.32255</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table 2**

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

<table>
<thead>
<tr>
<th>Tree</th>
<th>ML*</th>
<th>MP</th>
<th>NJ-JC</th>
<th>NJ-K2P</th>
<th>NJ-LogDet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A, B, (C, D))</td>
<td>−2,486.40247</td>
<td>282</td>
<td>0.31591</td>
<td>0.32010</td>
<td>0.32255</td>
</tr>
<tr>
<td>(A, C, (B, D))</td>
<td>−2,500.71061</td>
<td>294</td>
<td>0.35009</td>
<td>0.35900</td>
<td>0.35478</td>
</tr>
<tr>
<td>(A, D, (B, C))</td>
<td>−2,485.78542</td>
<td>280</td>
<td>0.31589</td>
<td>0.32001</td>
<td>0.32373</td>
</tr>
</tbody>
</table>

**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-
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-informa-

<table>
<thead>
<tr>
<th>Tree</th>
<th>GG98⁺</th>
<th>GG98⁻¹⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>((A, B), (C, D))</td>
<td>-2.460.608305</td>
<td>-2.430.073634</td>
</tr>
<tr>
<td>((A, D), (B, C))</td>
<td>-2.460.299516</td>
<td>-2.431.704680</td>
</tr>
<tr>
<td>((A, C), (D, B))</td>
<td>-2.474.811480</td>
<td>-2.434.132779</td>
</tr>
<tr>
<td>(A, (B, (C, D)))</td>
<td>-2.457.770740</td>
<td>-2.428.829561</td>
</tr>
<tr>
<td>(B, (A, (C, D)))</td>
<td>-2.460.911565</td>
<td>-2.430.073634</td>
</tr>
<tr>
<td>(C, (D, (A, B)))</td>
<td>-2.460.000532</td>
<td>-2.428.909027</td>
</tr>
<tr>
<td>(D, (C, (A, B)))</td>
<td>-2.459.423790</td>
<td>-2.429.815718</td>
</tr>
<tr>
<td>(A, (C, (B, D)))</td>
<td>-2.472.723745</td>
<td>-2.433.514116</td>
</tr>
<tr>
<td>(C, (A, (B, D)))</td>
<td>-2.473.956835</td>
<td>-2.432.968496</td>
</tr>
<tr>
<td>(B, (D, (A, C)))</td>
<td>-2.474.946236</td>
<td>-2.434.089933</td>
</tr>
<tr>
<td>(D, (A, (C, B)))</td>
<td>-2.473.359449</td>
<td>-2.434.094273</td>
</tr>
<tr>
<td>(A, (D, (B, C)))</td>
<td>-2.459.887635</td>
<td>-2.431.355763</td>
</tr>
<tr>
<td>(D, (A, (B, C)))</td>
<td>-2.460.380030</td>
<td>-2.431.765262</td>
</tr>
<tr>
<td>(B, (C, (A, D)))</td>
<td>-2.464.581165</td>
<td>-2.432.304177</td>
</tr>
<tr>
<td>(C, (B, (A, D)))</td>
<td>-2.463.186408</td>
<td>-2.430.966041</td>
</tr>
</tbody>
</table>

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

⁻ GG98 model with discrete gamma rate heterogeneity (four rate categories).
⁻⁻ Maximum-likelihood topology for both models.
These observations indicate the presence of as-yet-un-
known 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 fac-
tors 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 likeli-
hood can correct for problems that are identified and
incorporated into substitution models but can be de-
ceived 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 com-
bination with site-to-site rate heterogeneity and tran-
sition/transversion bias can cause problems even at bi-
ologically 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 tran-
sition/transversion bias and degree of rate heterogeneity
are not at all extreme. LogDet/paralinear distances pro-
vide 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 evolution-
ary factor, it is nevertheless beneficial to thoroughly an-
alyze sequence data sets in the search for good examples
of the effects of evolutionary factors representing po-
tential pitfalls for phylogeny methods. Equally impor-
tant is the search for new evolutionary factors. It is only
when such evolutionary factors as site-to-site rate het-
erogeneity, transition/transversion bias, evolutionary de-
pendence 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, Mi-
ichael Steel, Michael Hendy, and David Penny for mak-
ing the data sets used in their 1994 paper freely available
to other researchers over the World Wide Web. Permis-
sion 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

<p>| Table 4 |
| Results of Simulations Based on Parameter Estimates Made Using the GG98-1 Model |</p>
<table>
<thead>
<tr>
<th>Optimality Criterion</th>
<th>Model</th>
<th>% Correcta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum parsimony</td>
<td>Equal weights</td>
<td>94.2</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>F84</td>
<td>96.2</td>
</tr>
<tr>
<td>Minimum evolution</td>
<td>JC69</td>
<td>92.4</td>
</tr>
<tr>
<td></td>
<td>K80</td>
<td>91.5</td>
</tr>
<tr>
<td></td>
<td>LogDet</td>
<td>89.8</td>
</tr>
</tbody>
</table>

a Out of 1,000 simulation replicates.
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

\[
P_{ij}(t) = \begin{cases} 
\pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\beta t} + \left( \frac{\Pi_i - \pi_i}{\Pi_j} \right) e^{-\beta t[1+\Pi_j(k-1)]} & i = j \\
\pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\beta t} - \left( \frac{\Pi_i - \pi_i}{\Pi_j} \right) e^{-\beta t[1+\Pi_j(k-1)]} & i \neq j, \text{ transition} \\
\pi_j(1 - e^{-\beta t}) & i \neq j, \text{ transversion,}
\end{cases}
\]

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

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

Thus, the composition of \(A\) after time \(t\) is

\[
\pi_A(t) = \pi_A(0) \left[ \pi_A + \pi_A \left( \frac{1}{\Pi_R} - 1 \right) e^{-\beta t} \right. \\
+ \left. \left( \frac{\pi_A}{\Pi_R} \right) e^{-\beta t[1+\Pi_R(k-1)]} \right] \\
+ \pi_G(0) \pi_A + \pi_A \left( \frac{1}{\Pi_R} - 1 \right) e^{-\beta t} \\
- \left( \frac{\pi_A}{\Pi_R} \right) e^{-\beta t[1+\Pi_R(k-1)]} \\
+ \pi_C(0) \pi_A(1 - e^{-\beta t}) + \pi_T(0) \pi_A(1 - e^{-\beta t}) \\
= \pi_A + \pi_A \left( \frac{\Pi_R - \pi_R(0)}{\Pi_R} \right) e^{-\beta t} \\
+ \left( \frac{\pi_A(0) \Pi_R - \pi_R(0) \pi_A}{\Pi_R} \right) e^{-\beta t[1+\Pi_R(k-1)]},
\]

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

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

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 \((1 + [\beta(1 + \Pi_i(k - 1))/\alpha])^{-\alpha}\) for \(e^{-\beta t[1+\Pi_i(k-1)]}\):


MASAMI HASEGAWA, reviewing editor

Accepted February 13, 2001