

## COMPOSITIONAL BIAS, CHARACTER-STATE BIAS, AND CHARACTER-STATE RECONSTRUCTION USING PARSIMONY

TIMOTHY M. COLLINS,<sup>1</sup> PETER H. WIMBERGER,<sup>1,2</sup> AND GAVIN J. P. NAYLOR<sup>1</sup>

<sup>1</sup>Department of Biology, University of Michigan, Ann Arbor, Michigan 48109-1048, USA

**Abstract.**—Compositional bias, the occurrence of the four bases in unequal proportions, is a common feature of nucleotide sequences. We analyzed patterns of character-state reconstruction using maximum parsimony in two empirical data sets exhibiting compositional bias. For each case in which the inferred reciprocal numbers of changes for a pair of bases differed substantially, the two bases also differed markedly in relative abundance and the asymmetry of reconstructed transformations favored changes from the common state to the rare state. In addition, the compositional biases of the inferred ancestral sequences were more extreme than that seen in the terminal taxa, having an excess of common states relative to the terminals. Both of these features suggested that patterns of character-state reconstruction might be systematically distorted when compositional bias results in unequal representation of character states within a character, a condition we term character-state bias. Character-state bias is essentially compositional bias within a character. Simulation studies showed that highly asymmetric patterns of character-state reconstruction can be produced in the face of an underlying symmetry of character-state transformations in the presence of compositional bias. Rates of change are also important. The asymmetry of transformations produced in the simulations matched the pattern found in empirical data sets, with transformations from the common state to the rare state being more abundant. Rare states tend to be autapomorphic, typically requiring a change to the rare state on a tree. Conversely, changes from rare states to common states are systematically lost. These results are significant for methods that rely on character-state reconstruction using maximum parsimony, for example, to develop weighting schemes for phylogenetic analysis or to study patterns of correlated character evolution. Techniques that rely on character-state reconstruction may often be compromised by the distorting influence of character-state bias. [Compositional bias; character-state bias; character-state reconstruction; parsimony.]

One of the primary advantages of maximum parsimony over other methods of phylogenetic inference is that it allows for the reconstruction of ancestral character states at internal nodes (Donoghue, 1989; Maddison and Maddison, 1992). Knowledge of ancestral character states is a prerequisite to the study of character evolution. For example, hypotheses regarding the origin of adaptations and correlated character evolution require reconstruction of ancestral character states (Coddington, 1988; Maddison, 1990). The efficacy of many methods currently used in evolutionary biology hinges on the assumption that ancestral character states can be correctly re-

constructed. In this paper, we show that maximum parsimony can result in systematically biased ancestral character-state reconstructions when the different states within a character are in unequal proportions. The purpose of this communication is to promote caution and judicious interpretation of ancestral character-state reconstructions when character states differ markedly in their frequencies for a class of characters. We present two empirical case studies using DNA sequence data in which maximum parsimony implies ancestral character states that are likely incorrect and demonstrate the effect with computer simulations.

### BACKGROUND

DNA sequences lend themselves well to studies investigating evolutionary dynamics

<sup>2</sup> Present address: Department of Biology, University of Puget Sound, Tacoma, Washington 98416, USA.

because all sites (evolutionary characters) have the same four possible character states: the bases adenine (A), guanine (G), cytosine (C), and thymine (T) (alignment gaps are a fifth possible state when insertions or deletions have occurred). Thus, observations can be pooled across a number of sites (e.g., third positions of codons), and specific types of substitutional changes (i.e., transitional changes, transversional changes, and the 12 possible base substitutions) can be contrasted. This capacity for pooling observations effectively provides a larger sample size with which to detect subtle patterns of character change that might otherwise go undetected if sites (characters) had to be analyzed individually. Information about the dynamics of DNA sequence change over evolutionary time has been accumulating rapidly, and the possibility exists that comparatively simple stochastic models may adequately describe evolutionary change, especially silent changes, such as the third positions of codons in protein-encoding genes, which frequently undergo substitutions that do not affect the resulting amino acid sequence. These models can assist us in choosing and weighting characters for use in phylogenetic analysis and in assessing the reliability of trees, rate determinations, and patterns of character evolution inferred from parsimony methods (Penny et al., 1990; Maddison and Maddison, 1992).

The wealth of information about the dynamics of DNA sequence evolution that has accumulated to date can be used as a guideline for computer simulations of DNA sequence evolution along preordained evolutionary paths. Such simulations are useful for investigating the performance properties of different phylogenetic inference methods. The fidelity with which the inference method reflects the evolution that was simulated can be used as a relative measure of performance, to the extent that the model is a reasonable facsimile of the relevant evolutionary processes. Furthermore, simulations can be run under a number of different conditions to determine the ranges over which different inference methods perform well and at which point they begin to break down.

#### *Compositional Bias and Character-State Bias*

One common feature of nucleotide sequences is compositional bias—the occurrence of the four bases A, G, C, and T in unequal proportions. The degree of compositional bias varies widely among genes and organisms. For example, the G + C content of third positions of codons from nuclear and mitochondrial protein-encoding genes in bacteria, vertebrates, and insects ranges from 4% to 98% (Bernardi et al., 1985; Ikemura, 1985; Jukes and Bhushan, 1986; Sueoka, 1988; Liu and Beckenbach, 1992). When taxa under study exhibit a similar pattern and degree of compositional bias, they are said to exhibit stationarity, also sometimes referred to as base compositional equilibrium (Saccone et al., 1989). Variation in compositional bias among taxa is known as deviation from stationarity. Nucleotide sequences from different species within a clade may thus display compositional bias but maintain stationarity. Deviations from stationarity, however, necessarily involve compositional bias. These two features, compositional bias and deviation from stationarity, may be referred to jointly as compositional effects.

The potentially confounding influence of compositional effects on phylogenetic reconstruction has been considered for maximum parsimony, evolutionary parsimony, maximum likelihood, and distance methods (Loomis and Smith, 1990; Penny et al., 1990; Sidow and Wilson, 1990, 1991; Lockhart et al., 1992; Forterre et al., 1993; Hasegawa and Hashimoto, 1993; Sogin et al., 1993; Steel et al., 1993). Compositional effects may also influence estimates of base substitution rate, levels of apparent saturation, and procedures that attempt to “correct” divergence estimates to account for multiple substitutions at a site (DeSalle et al., 1987; Saccone et al., 1989, 1990; Kondo et al., 1993). The majority of these studies have focused on the effects of deviations from stationarity on such topics as inferred tree topologies and estimates of rates. In the present study, we consider the effect of compositional bias on character-state reconstruction using maximum par-

simony. We consider those cases in which unequal overall representation of the four bases in sequences (compositional bias) results in unequal representation of character states within a character, a condition we term character-state bias. Compositional bias will typically, though not necessarily, lead to character-state bias. To the extent that character states are more or less uniformly distributed among characters, compositional bias should result in character-state bias. If, however, character states are not uniformly distributed among characters, because of selection or some other type of constraint, compositional bias need not result in character-state bias. For example, a character state could be rare over all sites in a series of sequences being compared but common at those sites at which it does occur.

#### COMPOSITIONAL BIAS AND CHARACTER-STATE RECONSTRUCTION: EMPIRICAL STUDIES

The genus *Nucella* comprises several species of predatory gastropods of rocky intertidal marine habitats. This genus diversified during the Neogene and Quaternary in the North Pacific. In a study of molecular systematics and evolution of *Nucella* (Collins et al., in press), a 718-base-pair (bp) portion of the mitochondrial cytochrome *b* gene was sequenced for eight ingroup and two outgroup taxa and subsequently used to infer a phylogeny (Fig. 1). The cytochrome *b* gene of each of these species exhibits a strong compositional bias, particularly at third positions of codons, where the combined percentage of G + C on the coding strand is only 12–25% (Fig. 1). When the average numbers of changes between states for third positions were inferred (Fig. 2), an interesting relationship between the changes inferred and compositional bias became apparent. For each case in which a substantial asymmetry existed between the reciprocal numbers of changes for a pair of bases, the asymmetry was in the direction that made change to the rare state more common (Fig. 2). For example, in the analysis in Figure 1, where A is common and G is rare, the average

inferred A → G changes outnumber G → A changes by more than 3:1 (Fig. 2). Figure 2 also shows that changes of G to other states that might then change to A are too few to account for this difference. If a DNA sequence is stationary within a clade, we expect that the overall number of A → G and G → A changes should be roughly equal. We can certainly imagine a pattern of character evolution that would result in the maintenance of base compositional equilibrium given highly asymmetric reciprocal numbers of changes (Fig. 3), but this pattern would not be reconstructed by maximum parsimony methods in the manner required by our inferred numbers of changes. In general, if A → G changes predominate over G → A changes, as is suggested by the character-state reconstructions in *Nucella*, we would expect the number of G's to be increasing. This prediction is difficult to reconcile with the strong compositional bias against G's in *Nucella* and with the fact that most of the ingroup species have either maintained a compositional bias similar to that found in the outgroup species or, in several cases, experienced a decline in the number of G's relative to the outgroup taxa (Fig. 1). Another way to approach this question is to look at the inferred base composition at the base of the *Nucella* clade. If the base composition of the basal node of the *Nucella* lineage is inferred using parsimony, a compositional bias more extreme (a lower percentage of G + C) than in either the ingroup or outgroup results, particularly if only unequivocal reconstructions are considered (Fig. 1). It is certainly possible, if perhaps unlikely, that the *Nucella* progenitor had a base composition more extreme than either its descendants or the outgroup species. However, there is currently no reason to believe that ancestors will typically have compositional biases that are more extreme than those of the majority of descendants and their sister taxa.

To see if (1) the relationship between compositional bias and the direction of asymmetry of transformations and (2) the predominance of common states in the reconstruction of ancestors were peculiar to

T	50.8	44.6	43.7	42.9	47.9	52.1	51.7	48.7	51.3	50.0
A	33.7	32.5	32.1	32.5	37.1	35.8	35.0	34.2	30.8	30.8
C	8.8	13.3	15.0	15.8	12.1	7.1	8.7	10.8	10.4	11.7
G	6.7	9.6	9.2	8.8	2.9	5.0	4.6	6.3	7.5	7.5

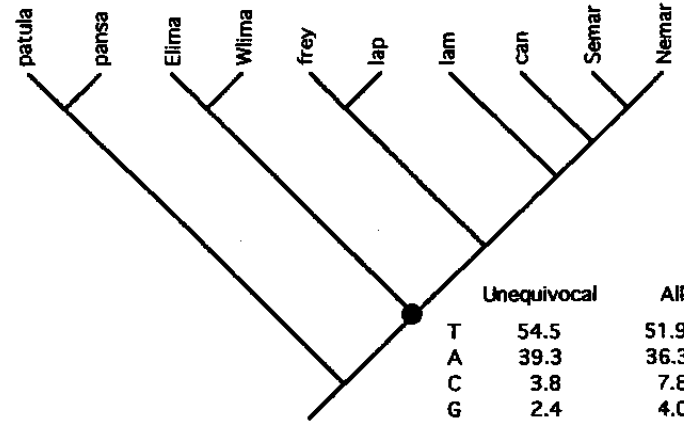


FIGURE 1. Most-parsimonious tree generated from 718 base pairs of gastropod mitochondrial cytochrome *b* sequence for species of *Nucella* and two outgroup species in the genus *Plicopurpura* (Collins et al., in press). The taxa (left to right) are *Plicopurpura patula*, *P. pansa*, *Nucella lima* (eastern and western forms), *N. freycineti*, *N. lapillus*, *N. lamellosa*, *N. canaliculata*, and *N. emarginata* (southern and northern forms). The percentage of each base at third positions is listed above each taxon. ● = the node at which the base composition of the progenitor of the *Nucella* species studied was inferred using the trace characters option of MacClade 3.0. Unequivocal character states at that node were tallied directly. For characters with equivocal reconstructions, the base composition was apportioned among the possible reconstructions. For example, if there were four most-parsimonious reconstructions (MPRs) and the node was C in three MPRs and T in one, the composition at that site was counted as 0.75 C and 0.25 T. The inferred base composition is reported for unequivocal reconstructions only and for all reconstructions.

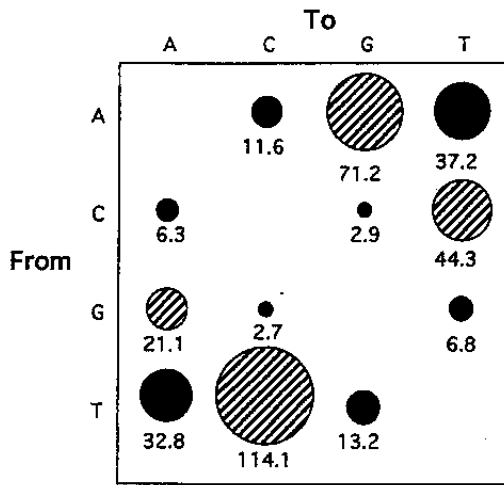


FIGURE 2. *Nucella* state changes chart. Average numbers of changes between states for third codon positions of *Nucella* cytochrome *b* data as estimated using the state changes and stasis option of MacClade 3.0. Bubble area is proportional to the inferred number of each type of change (number also shown below each bubble). ◐ = transitions; ● = transversions.

*Nucella* cytochrome *b*, we analyzed a mitochondrial cytochrome *b* data set from the pecoran ruminants (Irwin et al., 1991). The average numbers of changes between states

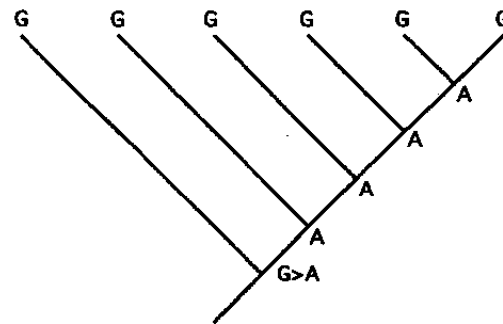


FIGURE 3. Example of stationarity in the face of asymmetry of transformations. For the character represented here, the character state is G in all of the terminal taxa. There is a transformation from G to A at the base of the ingroup followed by independent reversals from A to G in each terminal taxon.

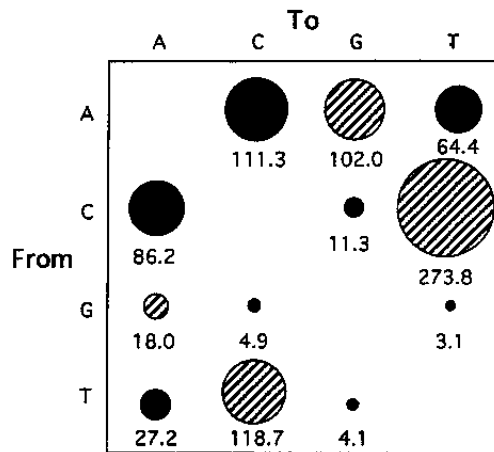


FIGURE 4. Pecoran state changes chart. Average numbers of changes between states for third codon positions of pecoran cytochrome *b* data as estimated using the state changes and stasis option of MacClade 3.0. Bubble area is proportional to the inferred number of each type of change (number also shown below each bubble). ○ = transitions; ● = transversions.

and the base composition of the pecoran progenitor were inferred on the parsimony tree presented by Irwin et al. (1991), although pecoran relationships are unsettled (Kraus and Miyamoto, 1991). Dolphin, camel, and

chevrotain sequences were used as outgroups. Once again, in each case where the reconstructed reciprocal numbers of changes differed substantially for a pair of bases, common-to-rare state changes predominated over rare-to-common changes (Fig. 4). For instance, average inferred  $A \rightarrow G$  changes were five times more common than  $G \rightarrow A$  changes, and  $C \rightarrow T$  changes outnumbered  $T \rightarrow C$  changes by more than 2:1, even though most of the ingroup members had fewer T's and G's than did the outgroup species (Fig. 4). The pattern of compositional bias differs between the pecoran and *Nucella* sequences. The *Nucella* sequences are A + T rich (Fig. 1), and the pecoran sequences are A + C rich (Fig. 5). In the *Nucella* data set, T's are more common than C's, and  $T \rightarrow C$  changes are the predominant pyrimidine transition; in the pecoran data set C's are more common than T's, and the  $C \rightarrow T$  changes are more abundant. When the base composition of the pecoran progenitor's DNA was inferred, a compositional bias more extreme than any found in the ingroup or outgroup resulted (Fig. 5).

These results demonstrate that the patterns found are not unique to the *Nucella*

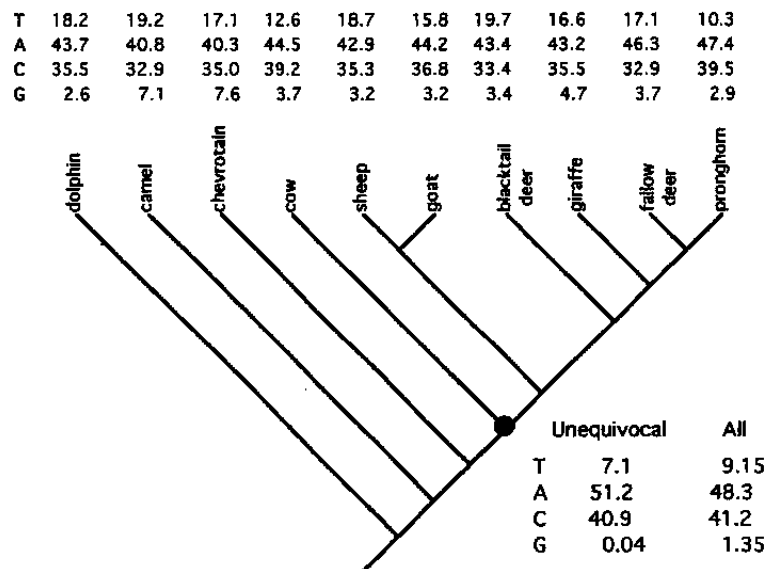


FIGURE 5. Pecoran tree from Irwin et al. (1991) with third codon position base composition of terminal taxa and base composition of pecoran progenitor inferred by methods as described for Figure 1.