

Step Matrices and the Interpretation of Homoplasy

RICHARD H. REE^{1,2} AND MICHAEL J. DONOGHUE

¹*Department of Organismic and Evolutionary Biology, Harvard University, 22 Divinity Avenue, Cambridge, Massachusetts 02138, USA*

Abstract.—Assumptions about the costs of character change, coded in the form of a step matrix, determine most-parsimonious inferences of character evolution on phylogenies. We present a graphical approach to exploring the relationship between cost assumptions and evolutionary inferences from character data. The number of gains and losses of a binary trait on a phylogeny can be plotted over a range of cost assumptions, to reveal the inflection point at which there is a switch from more gains to more losses and the point at which all changes are inferred to be in one direction or the other. Phylogenetic structure in the data, the tree shape, and the relative frequency of states among the taxa influence the shape of such graphs and complicate the interpretation of possible permutation-based tests for directionality of change. The costs at which the most-parsimonious state of each internal node switches from one state to another can also be quantified by iterative ancestral-state reconstruction over a range of costs. This procedure helps identify the most robust inferences of change in each direction, which should be of use in designing comparative studies. [Ancestral states; character evolution; homoplasy; parsimony; phylogenetic inference.]

The identification and interpretation of homoplasy is a subject of increasing interest, due in large part to advances in phylogenetic analysis (Sanderson and Hufford, 1996). Of particular relevance to the study of homoplasy are inferences of ancestral states on a phylogenetic tree and the character transitions inferred from them. The most commonly used criterion for optimizing ancestral states on phylogenies is maximum parsimony (Maddison, 1994). Parsimony is appealing in its simplicity, but the acceptance of hypotheses based upon it depends on background assumptions about the evolutionary process (e.g., Sober, 1988; Maddison, 1994).

Generalized parsimony algorithms for ancestral-state optimization (e.g., Sankoff and Cedergren, 1983; Swofford and Maddison, 1992) quantify such assumptions in the form of a matrix of costs associated with each type of character-state change (e.g., gain or loss of a trait); this is commonly known as a step matrix (Maddison and Maddison, 1992). The step matrix determines the most-parsimonious assignment of ancestral states on a given tree and therefore bears on such hypotheses as whether a trait was independently gained or secondarily lost. Step-matrix costs represent an assumption

about the extent to which the evolution of a particular character exhibits directionality, i.e., whether changes in one direction are favored over changes in the other. In this way they are analogous to probabilities of change in evolutionary models used in maximum likelihood estimation.

The assignment of equal costs to character-state changes, while often the default approach in studies of character evolution, is as much an assumption about evolution as are step matrices with unequal costs (Swofford and Maddison, 1992). While equal costs of change may be viewed as an appropriate starting point in the absence of any other information about a character, one might wish to incorporate unequal costs in light of evidence from various sources, such as studies of population genetics or ontogeny. For example, although bilaterally symmetrical flowers are generally assumed to have arisen many times independently from radially symmetrical ancestral forms, developmental genetic studies suggest that it is relatively easy for bilaterally symmetrical flowers to revert to radial symmetry (Coen and Nugent, 1994). Therefore, bilateral symmetry may have arisen deeper in angiosperm phylogeny than is generally assumed, and reversals to radial symmetry might have occurred more often. An approach to character analysis is needed that allows alternative assumptions about the

²Corresponding author. E-mail: rree@oeb.harvard.edu

directionality of evolutionary change to be assessed quantitatively. This issue has been explored in recent studies of the evolution of angiosperm breeding systems (Kohn et al., 1996), xylem (Bass and Wheeler, 1996), and dichromatic plumage in ducks (Omland, 1997).

Studies of the sensitivity of various aspects of phylogenetic inference have clarified the conditions under which evolutionary inferences may be more or less robust to underlying assumptions. While considerable attention has been paid to the accuracy of ancestral-state reconstructions under parsimony (Collins et al., 1994; Maddison, 1994; Maddison, 1995; Zhang and Kumar, 1997), the relationship between inferences of character evolution and step matrices has not been explored systematically.

SENSITIVITY OF INFERENCES TO ASSUMED COSTS

Here we develop a graphical approach to the sensitivity of inferred changes to cost assumptions, using a hypothetical example involving 64 taxa (Fig. 1), of which 25 exhibit one condition (state 1, solid branches) and 39 exhibit another (state 0, open branches). The number of independent gains and losses of state 1 can be inferred from a most-parsimonious optimization of ancestral states on the phylogeny shown in Figure 1. Based on the assumption that the cost of gaining state 1 (C_G) is equal to the cost of losing state 1 (C_L), there are more inferred gains (5) than losses (2).

Given the initial assumption that $C_G = C_L$, one might wonder how large the ratio C_G/C_L could get before an internal node that had been initially optimized to possess state 0 would become equivocal, or would be more parsimoniously assigned state 1. By iterative ancestral-state reconstruction over a range of values of C_G/C_L , the sensitivity of the optimization in Figure 1 can be quantified as the value C_{Gmax} , which represents an upper bound of the step matrix assumptions, beyond which the hypothesis that state 0 is the best assignment to a particular node is no longer supported by the data. Likewise, for nodes initially optimized as state 1, C_{Lmax} represents the maximum value of C_L/C_G for

which that optimization is the most parsimonious. The numbers mapped to internal nodes in Figure 1 represent values of C_{Gmax} and C_{Lmax} , and the double-headed arrows indicate the branches/nodes over which the indicated values apply.

Shifting attention to the overall pattern of change in the numbers of inferred gains and losses on the tree, we can safely assume that step matrices that assign a higher cost to losses (i.e., $C_G < C_L$) will tend to infer more gains, whereas step matrices that assign a higher cost to gains ($C_G > C_L$) will tend to infer more losses. This is confirmed in Figure 2a, which shows the results of iterative ancestral-state optimization over a range of step-matrix values for the phylogeny and character data from Figure 1. At each iteration, the cost of change in one direction (gain or loss) was changed by increments of 0.1, while the cost of change in the other direction was held constant at 1, and the number of unambiguous changes in each direction was determined from the most-parsimonious ancestral-state reconstruction. The number of inferred gains increases and the number of inferred losses decreases as C_G decreases relative to C_L , and vice versa. Although change in the number of gains or losses over a range of C_G/C_L values is a step function, in Figures 2 and 3 the steps are connected into a continuous line to help visualize and compare patterns of change in the numbers of gains and losses.

One noteworthy feature of this type of graph is that it identifies the range of values on the x -axis over which more gains are inferred than losses and more losses are inferred than gains. In some cases, an intermediate range covering one or more real values of C_G/C_L will yield an equal number of gains and losses. In other cases, such as that shown in Figure 2a, there are no real values at which the numbers of gains and losses are equal, and the point on the x -axis where the relative frequency of gains versus losses is reversed can be termed an inflection point. Hypotheses regarding the relative frequency of gains versus losses of a trait that are based on most-parsimonious ancestral-state reconstructions are therefore contingent upon which side of the inflection

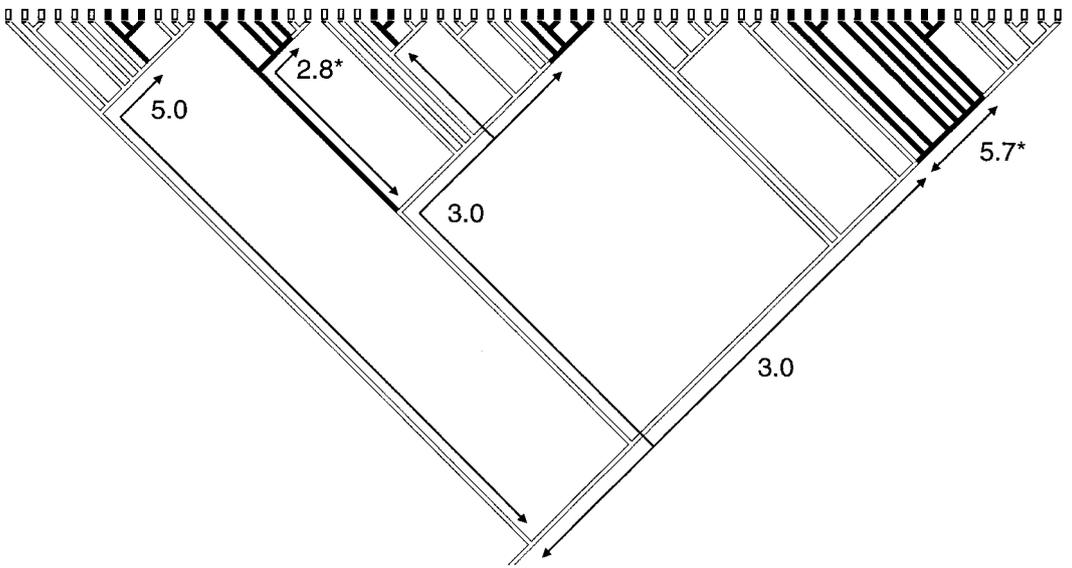


FIGURE 1. Hypothetical phylogeny on which inferred evolutionary changes in a binary character have been optimized by using parsimony, under the assumption that $C_G = C_L$. Based on this optimization, state 1 (solid branches) has evolved independently five times and has been lost twice. Values of C_{Gmax} and C_{Lmax} (marked with an asterisk) are indicated for internal nodes; double-headed arrows trace the branches over which the indicated values hold.

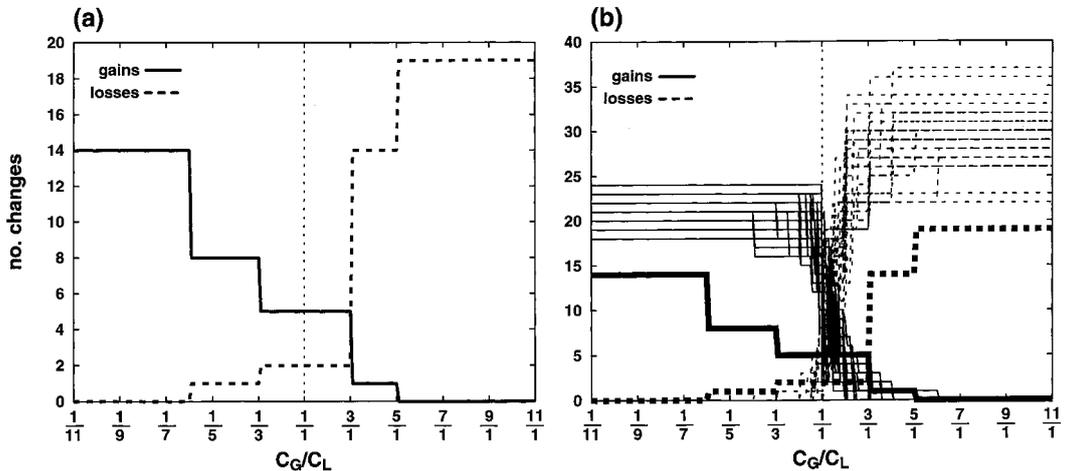


FIGURE 2. Cost-change graphs for the phylogeny in Figure 1. (a) The relationship between the frequency of unambiguous gains and losses inferred by using parsimony and the ratio of the cost of gains (C_G) and the cost of losses (C_L). As C_G increases relative to C_L , the number of gains decreases and the number of losses increases, and vice versa. If $C_G/C_L \leq 3$, more gains than losses are inferred; however, if $C_G/C_L > 3$, more losses than gains are inferred. The inflection point indicates where the lines representing gains and losses cross (see text). (b) Data from Figure 1 (bold) compared with 50 random permutations of the same data. Phylogenetic structure in the original data is evident in that, at each point on the x-axis, fewer changes are inferred than are inferred from the permuted data.

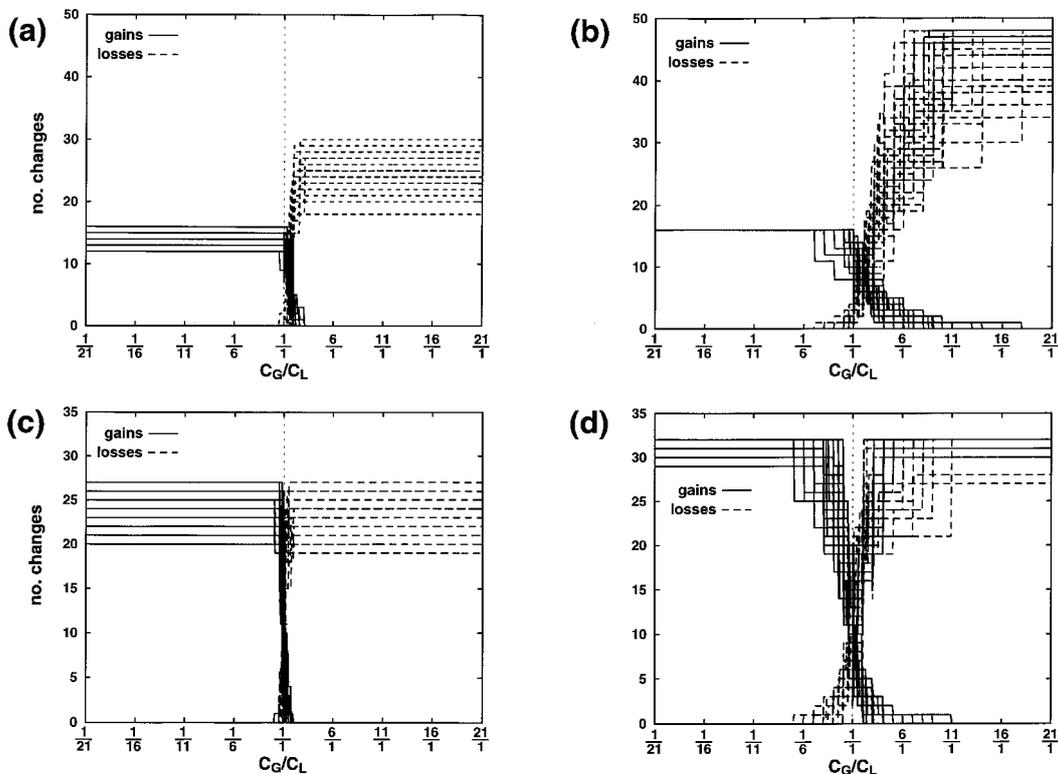


FIGURE 3. The effect of tree shape and tip state frequency on the relationship between inferred character changes and step-matrix assumptions. In each graph, the inferred changes from 50 randomly permuted binary characters are plotted on a 64-taxon tree. (a) Balanced tree, 16 of the 64 terminal taxa with state 1. (b) Pectinate tree, 16 of the 64 terminal taxa with state 1. (c) Balanced tree, equal numbers of terminal taxa with each state. (d) Pectinate tree, equal numbers of terminal taxa with each state.

point the step-matrix costs are assumed to lie. Moving away from $C_G = C_L$, change in the number of gains and losses will occur at values on the x -axis that correspond to the values of C_{Gmax} and C_{Lmax} shown mapped to internal nodes in Figure 1. Eventually, all inferred changes will be in the same direction, either all gains or all losses. In this example, fixation for change in one direction is achieved more rapidly on the right-hand side of the graph than on the left.

Graphs like the one shown in Figure 2a, which we will refer to as cost-change graphs, can be shown to vary in a predictable way with phylogenetic structure (the tendency for related taxa to share the same state). Phylogenetic structure tends to reduce the number of inferred changes as compared

with randomly permuting the states among the taxa, which effectively removes structure. Under parsimony, and assuming that no cost is incurred to simply maintain a state, a clade that is monomorphic for a state will be assigned that condition at all of its internal nodes, irrespective of the step matrix. In Figure 2b, the data from the example phylogeny in Figure 1, which clearly exhibit some degree of structure, are plotted in comparison with graphs of random permutations of the same data, showing that at all points along the x -axis the changes inferred from the autocorrelated data are overall fewer.

The effect of tree shape and the ratio of tip states on cost-change graphs is shown in Figure 3. Balanced, or symmetrical, dichoto-

mous trees (Figs. 3a, 3c) tend to be more sensitive to step-matrix perturbation than pectinate trees (Figs. 3b, 3d), in that graphs produced from the former tend to show steeper descent to fixation of only gains or only losses than do the latter. This happens because the inferred gains that are most recalcitrant to change in C_G are nested higher within clades that are otherwise monomorphic for state 0, and vice versa. Pectinate trees are less recalcitrant to this change because both the average and the maximum possible number of internodes between the root and a tip is greater for pectinate trees than for balanced trees (for 64 taxa, the average is 32.5 internodes for pectinate trees and 6 for balanced trees).

The ratio of tip states also has predictable effects on aspects of the cost-change graph, such as the horizontal position of the inflection point. If the frequencies of the two states are equal, the inflection point will tend to occur where $C_G = C_L$ (Figs. 3c, 3d). However, if one state is rarer than the other (Figs. 3a, 3b), the first is more likely to be reconstructed as being derived on the tree (Collins et al., 1994), and consequently the inflection point of the data will tend to be horizontally displaced with respect to the point on the x -axis where $C_G = C_L$. If state 0 is rare relative to state 1, the inflection point will be displaced to the left; if state 1 is rare relative to state 0, displacement will be to the right.

The symmetry of cost-change graphs about the inflection point is also affected by the ratio of tip states. Step-matrix graphs become increasingly asymmetrical as the frequencies of the two states become more unequal (compare Figs. 3a, 3b and Figs. 3c, 3d). This is because the frequency of a character state provides an upper limit to the number of times that the trait can be inferred to have evolved. The impact of the relative frequency of character states on ancestral-state optimization highlights the importance of taxon sampling in comparative studies, as different sampling strategies may yield data sets with different proportions of character states.

C source code for the computer program used to iteratively reconstruct ancestral states, graph the relationship between num-

bers of gains and losses and C_G/C_L for real and permuted data, and find $C_{G_{\max}}$ and $C_{L_{\max}}$ values for internal nodes is available from one of us (R.H.R.: rree@oeb.harvard.edu), and from the Society of Systematic Biologists home page (<http://www.utexas.edu/depts/systbiol>).

DISCUSSION

Methods for exploring the relationship between step matrices and ancestral-state reconstructions are presented here in the hope that they will allow relevant sensitivity analyses to be more readily implemented in studies of character evolution, in the form of cost-change graphs and values of $C_{G_{\max}}$ and $C_{L_{\max}}$ for internal nodes on phylogenies. These methods obviously do not estimate the "true" transition costs or probabilities of change underlying the evolution of a particular trait, but this is not the objective; in fact, it is probably more realistic to assume that these parameters have varied throughout the evolution of a group as a result of changes in factors such as character burden (Riedl, 1978; Donoghue, 1989). Graphing the number of gains and losses inferred by parsimony over a range of assumptions, and quantifying the costs of change at which internal nodes switch from one ancestral-state reconstruction to the other, are simply means of assessing the sensitivity of hypotheses of character evolution that are based on single ancestral-state reconstructions. In principle, such sensitivity analyses could be extended to multistate characters, but for characters with more than two states, analogous graphs would require dimensions equal to the number of states and would be more difficult to construct and interpret. It should also be possible to develop similar procedures for continuously varying characters. A similar approach might be extended to sets of characters to quantify, for example, the sensitivity of inferred branch lengths to changes in cost assumptions.

We anticipate that the most important application of the methods described here will be in designing studies of evolutionary change between character states. Quantifying the robustness of inferred changes on

a phylogeny will help identify appropriate study taxa if the objective is to investigate the mechanisms underlying a particular direction of change. Suppose, for example, that you are a developmental geneticist interested in understanding shifts from bilaterally symmetrical to radially symmetrical flowers. You would like to focus your studies on the species with radially symmetrical flowers that are most certain to have been derived from a bilaterally symmetrical ancestor, and compare such species with their closest bilaterally symmetrical relatives. In this case the comparison of $C_{L_{\max}}$ values will facilitate the choice of study organisms. Those shifts marked by especially high values of $C_{L_{\max}}$ are most robust to changes in the step matrix, which means that the inferred direction of evolution (loss of bilateral symmetry) cannot easily be reversed by differentially weighting the costs of change. You can therefore be relatively confident that you are spending your time and money studying the kind of transition you are actually interested in. Those shifts marked by lower values of $C_{L_{\max}}$ represent cases in which the direction of evolutionary change is not as strongly supported by the phylogeny and might be reversed by only slight changes in assumptions. Here, one can be less certain that the organisms being studied actually represent lineages in which radial symmetry evolved from bilateral symmetry.

Another potential use of step-matrix sensitivity analyses that deserves attention is as a means of testing hypotheses of character evolution by a permutation procedure. Permutation tests have been developed for other aspects of phylogenetic analysis, e.g., monophyly (Faith, 1991) and structure (Faith and Cranston, 1991; Alroy, 1994), and one might suppose that such tests could be used to detect trends in character evolution. In comparing cost-change graphs of real character data with random permutations (e.g., as in Fig. 2b), it occurred to us that some aspect of the graph of the original data (or the value of $C_{G_{\max}}$ at a particular node) might be useful as a nonparametric test statistic for detecting significant deviation from random expectation. For example, one could compare the horizontal posi-

tion of the inflection point of a character (i.e., the point on the x -axis beyond which the hypothesis that more gains have occurred than losses, or vice versa, is no longer supported by a most-parsimonious reconstruction of ancestral states from the data) with a null distribution of inflection points generated by permutation. One has to permute the data because the null expectation for the position of the inflection point could be to the right or left of where $C_G/C_L = 1$, depending on the relative frequencies of character states in the data. If the real inflection point were to fall in the right-hand 5% tail of the null distribution, the value of C_G/C_L required to infer as many or more losses than gains would be greater than expected if the data were randomly permuted. A significant result would indicate that inferences of gains are more robust to changing the step-matrix costs (because they are more highly nested within clades exhibiting the opposite state) than are the inferences of losses. This might be interpreted to mean that the data are biased in favor of inferring more gains than losses. An alternative permutation-based test could use C_G as a test statistic, and compare it with a null distribution of $C_{G_{\max}}$ values. A value of C_G that is significantly higher than expected by chance alone for a particular node would support the assignment of state 0 to that node.

Such permutation tests for detecting trends in character evolution are, however, problematical for two reasons. First, although the behavior of cost-change graphs can be predicted with respect to particular factors in isolation (structure, tree shape, and the frequencies of tip states), the interactions among these factors are likely to be complex for any particular data set. Focusing on a single aspect of a cost-change graph (or $C_{G_{\max}}$ value) for statistical comparison with a null distribution would require an understanding of how the interaction of factors might affect the outcome, and therefore care must be taken in the development and application of such tests.

A second, more general, problem concerns the use of permutation as a means of generating a null distribution. The relationship between permutation and a null hy-

pothesis for the process of character change is unclear, so nonparametric statistical tests that rely on permutation-based null distributions should be treated cautiously. An alternative approach is to use maximum-likelihood methods, in which a probabilistic model of character change is made explicit, and one or more parameters of that model are estimated in such a way that the likelihood of observing the data is maximized (Edwards, 1992). Evolutionary hypotheses about the model parameters (e.g., that gains are more probable than losses) can then be tested statistically with likelihood ratios (Huelsenbeck and Rannala, 1997). In general, we believe that likelihood tests, along the lines of those developed by Sanderson (1993), Milligan (1994), and Pagel (1994), are more promising for studies of directionality than the parsimony-based permutation tests are.

ACKNOWLEDGMENTS

We thank David Baum, Brian Farrell, and participants in Biology 216 (Topics in Plant Evolution) for helpful discussion of these issues.

REFERENCES

- ALROY, J. 1994. Four permutation tests for the presence of phylogenetic structure. *Syst. Biol.* 43:430–437.
- BASS, P., AND E. A. WHEELER. 1996. Parallelism and reversibility in xylem evolution: A review. *Int. Assoc. of Wood Anat. J.* 17:351–364.
- COEN, E. S., AND J. M. NUGENT. 1994. Evolution of flowers and inflorescences. *Development (Suppl.)*: 107–116.
- COLLINS, T. M., P. H. WIMBERGER, AND G. P. NAYLOR. 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* 43:482–496.
- DONOGHUE, M. J. 1989. Phylogenies and the analysis of evolutionary sequences, with examples from seed plants. *Evolution* 43:1137–1156.
- EDWARDS, A. W. F. 1992. *Likelihood*. Johns Hopkins Univ. Press, Baltimore, Maryland.
- FAITH, D. P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Syst. Zool.* 40:366–375.
- FAITH, D., AND P. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? Permutation tests for cladistic structure. *Cladistics* 7:1–28.
- HUELSENBECK, J. P., AND B. RANNALA. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* 276:227–232.
- KOHN, J. R., S. W. GRAHAM, B. MORTON, J. DOYLE, AND S. C. H. BARRETT. 1996. Reconstruction of the evolution of reproductive characters in Pontederiaceae using phylogenetic evidence from chloroplast DNA restriction site variation. *Evolution* 50:1454–1469.
- MADDISON, D. R. 1994. Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. *Annu. Rev. Entomol.* 39:267–292.
- MADDISON, W. P. 1995. Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Syst. Biol.* 44:474–481.
- MADDISON, W. P., AND D. R. MADDISON. 1992. *MacClade: Interactive analysis of phylogeny and character evolution*. Sinauer Associates, Sunderland, Massachusetts.
- MILLIGAN, B. G. 1994. Estimating evolutionary rates for discrete characters. Pages 299–311 in *Models in phylogeny reconstruction (Systematics Association Special Volume 52; R. W. Scotland, D. J. Siebert, and D. M. Williams, eds.)*. Clarendon Press, Oxford, England.
- OMLAND, K. E. 1997. Examining two standard assumptions of ancestral reconstructions: Repeated loss of dichromatism in dabbling ducks (Anatini). *Evolution* 51:1636–1646.
- PAGEL, M. D. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond.* 255:37–45.
- RIEDL, R. 1978. *Order in living organisms*. Wiley, New York.
- SANDERSON, M. J. 1993. Reversibility in evolution: A maximum likelihood approach to character gain/loss bias in phylogenies. *Evolution* 47:236–252.
- SANDERSON, M. J., AND L. HUFFORD (eds.). 1996. *Homoplasy: The recurrence of similarity in evolution*. Academic Press, San Diego, California.
- SANKOFF, D., AND R. J. CEDERGREEN. 1983. Simultaneous comparison of three or more sequences related by a tree. Pages 2530–263 in *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison (D. Sankoff and J. B. Kruskal, eds.)*. Addison-Wesley, Reading, Massachusetts.
- SOBER, E. 1988. *Reconstructing the past: Parsimony, evolution, and inference*. The MIT Press, Cambridge, Massachusetts.
- SWOFFORD, D. L., AND W. P. MADDISON. 1992. Parsimony, character-state reconstructions, and evolutionary inferences. Pages 186–223 in *Systematics, historical ecology, and North American freshwater fishes (R. J. Mayden, ed.)*. Stanford Univ. Press, Palo Alto, California.
- ZHANG, J., AND S. KUMAR. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* 14:527–536.

Received 22 July 1997; accepted 15 February 1998

Associate Editor: D. Cannatella