eds., *Vicariance Biogeography: A Critique*, pp. 446–489. New York: Columbia University Press.
Platnick, N. I. and G. Nelson. 1978. A method of analysis for historical biogeography. *Syst. Zool.* 27:1–16.
Scudder, S. H. 1892. Some insects of special interest from Florissant, Colorado and other points in the territories of Colorado and Utah. *Bull. U.S. Geol. Surv.* 93: 25 pp., 3 pls.
Townsend, C. H. T. 1938. Five new genera of fossil Oestromuscaria (Diptera). *Entomol. News* 49:166–167.
Vermeij, G. 1989. Geographical restriction as a guide to the causes of extinction: the case of the cold northern oceans during the Neogene. *Paleobiology* 15:335–356.
Wille, A. 1977. A general review of the fossil stingless bees. *Rev. Biol. Trop.* 25:43–46.
Woodley, N. E. 1986. Parhadrestiinae, a new subfamily for *Parhadrestia* James and *Cretaceogaster* Teskey (Diptera: Stratiomyidae). *Syst. Entomol.* 11:377–387.

7 : Extinction, Sampling, and Molecular Phylogenetics

Ward C. Wheeler

Abstract. Extinction, whether natural or artificial, removes taxa from phylogenetic consideration. Here the effects of extinction on molecular data are examined through computer simulations. An analysis of variance (ANOVA) analysis is performed to examine the effects of several variables on the accuracy and resolution of phylogenetic reconstruction. Overall, the number of taxa used in cladogram construction is the most important factor in cladogram accuracy and the number of characters is most responsible for its resolution.

Extinction poses a dilemma for all systematics; molecular studies are no exception. For molecular systematists, the question is fundamental: Do taxa lost through extinction affect our interpretation of the relationships among living organisms? Using morphology, Gardiner (1982), following Patterson (1981), assumed in his study of amniotes that they do not. Gardiner constructed a phylogeny based solely on extant taxa, erecting the group Homeothermia, which unites birds and mammals. However, Gauthier et al. (1988) demonstrated persuasively that the inclusion of fossil taxa can greatly affect our phylogenetic understanding of these same organisms. In their

analysis, the inclusion of fossil synapsids caused the position of mammals to shift relative to lepidosaurs and chelonians (turtles). The influence of fossil taxa in such realignments was further emphasized by Donoghue et al. (1989). With very few exceptions, fossil organisms cannot be employed in molecular studies. In the construction of phylogenies can we compensate for this deficit in fossil taxa by turning instead to the wealth of molecular characters available from extant organisms?

The limitations imposed by natural extinction on molecular studies is further compounded by a form of artificial extinction, in which systematists draw on characters from a single taxon to represent a large and diverse group (Wheeler 1989). In these cases, the characters of that exemplar are assumed to reflect the ground plan of the entire group, but the establishment of this ground plan is impossible, throwing doubt on the synapomorphies linking higher taxa.

Analysis of phylogenetic relationships can also be hampered by missing data. In morphologic studies, the poor quality of a specimen or a condition of extreme apomorphy may render characters unobservable. In molecular systematics, on the other hand, the type of missing data encountered most often is a gap or gaps in aligned sequences. As with the loss of taxa through extinction, molecular systematists propose that the problem of missing data can be overcome by gathering more data.

Both missing characters and extinct taxa dilute the pool of data available to investigate phylogenetic patterns. In the first case, characters are lost, and in the second, taxa are lost, but they amount to the same thing: a decrease in the sample of phylogenetic information. Computer simulation is used here to examine the effects of missing characters and taxa on the efficacy of phylogenetic reconstruction based on molecular sequence data.

Simulations

There were three parts to the simulation procedure: the generation of the sequences, the culling of taxa, and the analysis of the phylogenetic results.

Evolution with both constant and variable rates of change was simulated. In the first model, the rate of anagenetic change was held constant throughout all lineages and for all characters. Similarly, the rate of cladogenesis was kept constant for all lineages. The second evolutionary model was more complex. Here the rates of evolution were determined uniquely for each character by a Poisson process. The probability of each lineage splitting was determined by another Poisson process. Thus, although the constant model simulated a clocklike evolution of characters and lineages, the variable model depicted a system with changing rates of character evolution and lineage splitting.

These two models of evolution were simulated using 25 different ratios of anagenesis versus cladogenesis. These ratios ranged from a splitting probability 50 times that of character change (hence a character has a 1 in 50 chance of transforming before the next lineage split) to a splitting probability one third that of character change. This rate variation yields scenarios wherein there is little (if any) character evolution between speciation events to those in which a great deal of evolution has occurred in the interval between lineage splits. At the most rapid rates of character change, the sequences approached randomness due to the extreme amounts of character change.

In both models, the nucleotide distribution was entirely symmetrical ($A=C=G=T=0.25$), as were the transformation probabilities among bases (all transformations equally likely). Sequences of three lengths—50, 100, and 250 bases—were generated.

All these permutations were repeated for varying proportions of missing data. There were four levels of deficit 0%, 5%, 10%, and 25%. Two procedures for choosing characters and taxa were employed. In the first, all characters were equally likely to become missing, whereas in the second, a Poisson distribution was used to establish a character's probability of becoming missing. All taxa had an equal chance of losing characters in both cases.

The purpose of the Poisson distribution was to add more variation to the system ($\mu=\sigma^2$, for Poisson). Under its influence, certain lineages and characters will evolve more rapidly than others. Similarly, certain base positions were more likely to be unobservable or missing. Both these situations seemed to represent the circumstances of evolution more realistically than simple homogeneous change due to the mechanical shortcomings of sequencing protocols, alignment procedures, and observed phenomena such as mutational "hot spots."

The second part of the simulation concerned the extinction process. Twelve sequences were generated for each data set using the preceding procedures, but only those with four monophyletic groups (with two or more taxa each) were retained. One representative sequence was chosen from each of these four groups at random, and a new, less diverse data set was created. This culling process was repeated 25 times for each of the data sets, to produce one full data set and 25 altered ones. Thus 26 data sets were generated for each combination of parameters: model of evolution, percent missing data, type of missing data, length of sequence, and ratio of anagenesis to cladogenesis.

These data sets were in turn subjected to Farris' HENNIG *86* (1988, Version 1.5) to determine the most parsimonious phylogenetic arrangement

of the taxa. In these runs, the characters were considered nonadditive (no transformation series among states was specified, hence no specific evolutionary model was assumed). The heuristic options mh*bb*, the most exhaustive of the heuristic tree searches involving the construction of multiple trees and branch-swapping, were used because implicit enumeration proved too time-consuming. When multiple, equally parsimonious solutions existed, the strict consensus method was used to summarize agreement among the most-favored hypotheses.

The results of these phylogenetic analyses were then tested for their similarity to the "correct" or known tree, as reported from the sequence-generating program. In the case of the four-taxa data sets, the tree produced either agreed completely with the arrangement of the four monophyletic groups in the original data or did not agree at all. The situation was more complex for the full, 12-taxa data set. Here concordance between the known tree and that produced was a matter of degree, with some members of the monophyletic groups correctly placed and some not. Thus agreement was assessed as the product of a series of fractions, representing the proportion of correctly placed members in each of the four monophyletic groups.

In the analysis, a further distinction is made between arrangements of taxa that are correctly placed and those that are not incorrectly placed or unresolved. An unresolved cladogram may not be "correct"—it neither conveys crucial information nor is necessarily misleading. Polytomous hypotheses make no statements about certain potential groups and avoid the "incorrect" placement of taxa. A fully resolved tree, on the other hand, can be utterly disinformative.

The level of resolution of the 12-taxa cladograms was defined as the number of resolved groups divided by the number of potentially resolved groups. For 12 taxa there are 10 nontrivial groups (all 12 always form a group); hence the index of resolution is the number of resolved groups divided by 10.

Results

Five factors affected the outcome of the phylogenetic trials: the number of taxa (12 or four), the length of the sequence modeled (number of characters), the type and degree of missing data, the rate of evolution, and the stochastic-model under which the sequences "evolved." All five conditions determined the percent of reconstructions that were not incorrect. Four of the factors were used to assess cladogram resolution; the number of taxa was not included in this assessment since the level of resolution only had meaning for the full data sets.

The results of the trials were evaluated through two multiway analyses of variance (ANOVA) procedures. In the first of these, each of the five factors was treated as an independent variable, with the success of the cladogram (percent not incorrect) as the dependent variable. The second analysis was a four-way procedure with the level of cladogram resolution as the dependent variable. Both of these ANOVAs were models without replication. Therefore interaction effects could not be separated from residual error. The ANOVA results are summarized in tables 7.1 and 7.2. Since the dependent values were percents, the data transformation $\arcsin\sqrt{\theta}$ was used to check normality. The additivity assumption of this model was examined with Tukey's test for additivity. For both ANOVA procedures—with and without data transformation—there was insignificant deviation from additivity (table 7.3).

As these tables show, four of the five factors were significant contributors to variation in cladogram success, as measured by fraction not incorrect: the

Table 7.1. Factors Affecting Cladogram Success (Percent not Incorrect)

		Unaltered Data		Transformed Data	
Factor	df	F	% variance	F	% variance
Number of taxa	1	11.9*	47.3	13.2	49.7
Sequence length	2	4.52†	17.9	4.57†	17.2
Missing data	6	0.22	0.89	0.28	1.07
Rate of evolution	24	2.70*	10.7	2.72*	10.2
Stochastic model	1	4.84†	19.9	4.77†	18.0

*Significant at $p<0.001$.
†Significant at $p<0.05$.

Table 7.2. Factors Affecting Cladogram Resolution

		Unaltered Data		Transformed Data	
Factor	df	F	% variance	F	% variance
Sequence length	2	22.6*	87.6	22.3*	85.8
Missing data	6	0.08	0.30	0.55	2.11
Rate of evolution	24	2.00†	7.76	1.94‡	7.47
Stochastic model	1	0.11	0.43	0.19	0.73

*Significant at $p<0.001$.
†Significant at $p<0.005$.
‡Significant at $p<0.01$.

Table 7.3. Examination of Additivity—Tukey Test

Analysis	F*	
	Unaltered Data	Transformed Data
Cladogram success	1.78×10^{-12}	8.07×10^{-12}
Cladogram resolution	5.19×10^{-6}	5.37×10^{-6}

* None of these values are significant at the $p = 0.05$ level.

number of taxa, sequence length, rate of evolution, and model of evolution. As for resolution, only sequence length and rate of evolution were significant. In neither case did missing data influence the phylogenetic outcome significantly.

The overwhelmingly important influence on the accuracy of the cladograms was the number of taxa, which accounted for 47% to 49% of the variance. The most important factor in the resolution of the cladograms was sequence length, responsible for 85 percent to 88 percent of the variance.

Four assumptions underlie any ANOVA procedure like this: normality, homoscedasticity, additivity, and independence. If the data do not conform to these prescriptions, the results of the analysis suffer accordingly. Although the method is somewhat robust to violations of the first two assumptions, both were checked through the use of the arcsin$\sqrt{\theta}$ transformation of the data. This transformation tends to normalize the distribution of percents and fractions (normality) while equalizing differences in variance (homoscedasticity). Since the results were unchanged by the application of arcsin$\sqrt{\theta}$, the likelihood of gross violation of either normality or homoscedasticity is low. The assumption of additivity—that the factors are linearly related to the outcome—was directly tested and upheld via the Tukey procedure. The final assumption of the method is independence (i.e., the factors are not inherently coupled in the determination of the cell values). Since the factors were varied externally by the experimental procedure, they are by their very nature independent. Thus the four assumptions appear to be met, and the variance in the data can be meaningfully partitioned among the factors mentioned earlier.

As shown in table 7.1, the number of taxa contributed the most to cladogram accuracy, accounting for almost one half of the variance. The second most important factor was the stochastic process or evolutionary model used to simulate the evolution of the sequences. Although the "constant" and "Poisson" model behaviors are similar, the Poisson model appears to reach a plateau, after which an increase in the number of characters fails to

have the same beneficial effect. In the constant, or "clock," model of change, the effect from an increase in data is more linear, increasing with increasing numbers of characters. This linearity may explain the view of some investigators that doubling the length of sequences will double the accuracy of phylogenetic reconstruction—even with only a few taxa (Steele et al. 1991). If evolution were truly clocklike, such an argument would be more compelling.

Approximately one-sixth of the variation in cladogram accuracy can be ascribed to the length of the sequences, or the number of characters. As the sequences become longer (fig. 7.1), the phylogenetic hypotheses become more reliable (but the effect is not as pronounced in the 12-taxon case). A distant fourth among the factors influencing cladogram accuracy is the rate of evolution, the ratio of anagenetic to cladogenetic change. This ratio determines the extent of character change likely to occur between lineage-splitting events in the "evolution" of the modeled sequences. There is a slight decrease

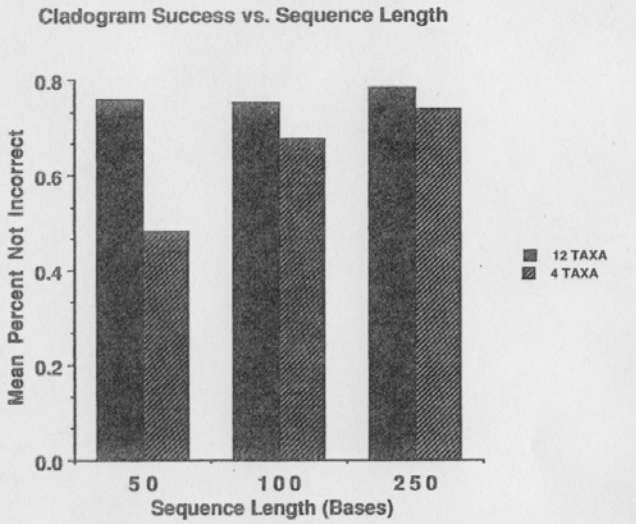

Fig. 7.1. Cladogram success (accuracy of reconstruction) versus sequence length. Although the increase in cladogram success for the complete data sets was not great, the culled data showed a large increment in success with increased numbers of characters.

in accuracy with very high and low rates of evolution. When rates are low, the sequences yield very little information about relationships among the taxa under study. When the rates are high, the sequences can approach randomness, containing little or no historical information. At both extremes, cladogram construction falters.

The second descriptor of cladogram behavior was resolution. The degree to which a cladogram was resolved offered a measure of its information content. If the scheme was completely resolved, whether correctly or incorrectly, it contributed a maximum of information on phylogenetic relationships. If it was completely unresolved, it yielded no information.

The factor of overwhelming importance for the resolution of the cladogram was sequence length. Almost 90% of the total variance was attributable to this one factor. Only the rate of evolution also contributed significantly, but to an order of magnitude less than the number of characters. The degree of resolution increased dramatically with longer sequences (fig. 7.2).

As with cladogram accuracy, the distorting effects of the rate of evolution

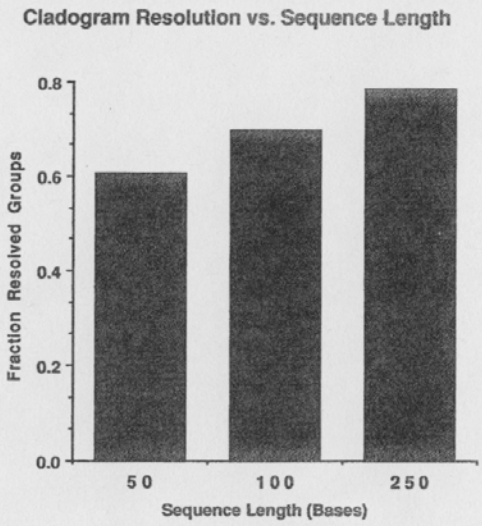

Fig. 7.2. Cladogram resolution versus sequence length. The resolution of cladograms increased markedly with increased number of characters.

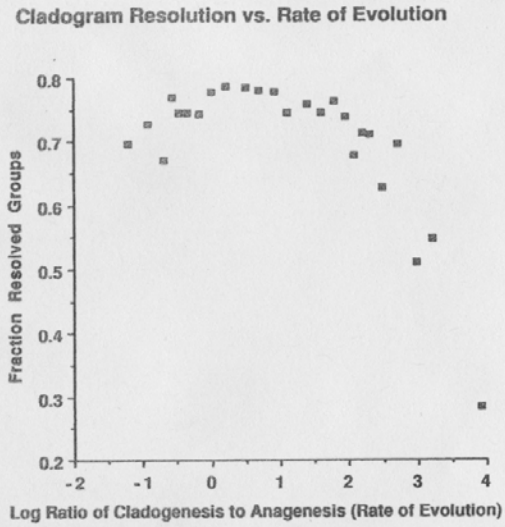

Fig. 7.3. Cladogram resolution versus rate of evolution. As with the accuracy of cladograms, the main effect of evolutionary rate comes in the extremely high and low values.

were most prominent at the very high and low ends of the range. When the rates were very low, cladograms were unresolved because of lack of sequence variation. At the highest rates, they were unresolved because of contradiction and confusion among the characters from the high number of hits (evolutionary events) at each position (fig. 7.3).

In neither of these analyses did missing data have a significant effect. In part, their lack of influence may stem from the way missing data were generated. There were two procedures by which missing data were introduced. The first assigned to each character an equal probability of becoming missing; thus the holes in the data matrix were completely random. In the second case, certain positions were more likely to go unobserved (missing) than others. This model reflected the circumstance in which certain characters are more difficult to observe, or are located in sequence areas of frequent insertion or deletion. But in no case were certain taxa (or specimens) more likely to contain missing data than others. The first two types of missing data

approximate situations that arise often in molecular sequence data. The case of missing data varying by taxon, however, echoes situations involving fossil or extremely derived taxa and was not examined here. Nonetheless, if the results of this simulation are credible, missing data may not present as much of a problem in DNA studies as they do in morphologic ones.

Another problem not examined here is the effect of evolutionary rate on initial homology statements. If the sequences have undergone multiple hits at a high proportion of their positions, it can be very difficult to align these positions. Here the homologies were assumed to be known. In reality, homologies would not be known with absolute certainty, allowing corresponding adverse effects on the reconstruction of phylogenetic arrangements from sequence data.

These results are described with a final caution concerning their robustness. As stated earlier, the evolutionary model used was a significant factor in the analysis. Yet in modeling evolution, we choose from a world of possible distributions without knowing which (if any) are correct. How robust are these conclusions, relying as they do on specific models? It is hoped that effects of number of taxa, sequence length, and rates of evolution that seem to hold up under both models will endure.

These caveats aside, it seems clear that the accuracy of a cladogram is most influenced by the number of taxa, the area in which the effect of extinction exerts greatest influence. When only a few taxa are used, or when only a small number are available, the accuracy of phylogenetic reconstruction suffers. The inclusion of ever-longer sequences will alleviate the problem somewhat; but more likely, it will merely increase the level of detail (i.e., resolution) offered by incorrect statements. Accordingly, the inclusion of the highest number of taxa possible, in the long run, will be the best way to ensure accurate results. Presumably, all the taxa are not required to confidently reconstruct the essential framework of phylogenetic relationships. Nonetheless, studies that rely on the analysis of less than 1% of pertinent diversity must always be viewed in light of the effects of extinction, whether artificial or natural.

ACKNOWLEDGMENTS

I would like to thank Ranhy Bang, Elise Broach, Michael Novacek, Paul Vrana, and Quentin Wheeler for helpful comments on this manuscript. This research was supported by the Alfred P. Sloan Foundation.

REFERENCES

Donoghue, M. J., J. Doyle, J. Gauthier, A. Kluge, and T. Rowe. 1989. The importance of fossils in phylogeny reconstruction. *Ann. Rev. Ecol. Syst.* 20:431–460.

Farris, S. J. 1988. HENNIG 86, version 1.5.

Gardiner, B. 1982. Tetrapod classification. *Zool. J. Linnean Soc.* 74:207–232.

Gauthier, J., A. G. Kluge, and T. Rowe. 1988. Amniote phylogeny and the importance of fossils. *Cladistics* 4:105–209.

Patterson, C. 1981. The significance of fossils in determining evolutionary relationships. *Ann. Rev. Ecol. Syst.* 12:195–223.

Steele, K. P., K. E. Holsinger, R. K. Jansen, and D. W. Taylor. 1991. Assessing the reliability of 5S rRNA sequence data for phylogenetic analysis in green plants. *Mol. Biol. Evol.* 8:240–248.

Wheeler, W. C. 1989. The systematics of insect ribosomal DNA. In B. Fernholm, K. Bremer, and H. Jörnvall, eds., *The Hierarchy of Life*, pp. 307–321. Amsterdam: Elsevier.