Evolution, 46(3), 1992, pp. 627–640

USING PHYLOGENETIC TREES TO STUDY SPECIATION AND EXTINCTION

JODY HEY
Rutgers University, Department of Biological Sciences, Nelson Laboratories, P.O. Box 1059, Piscataway, NJ 08855 USA

Abstract.—One tool in the study of the forces that determine species diversity is the null, or simple, model. The fit of predictions to observations, good or bad, leads to a useful paradigm or to knowledge of forces not accounted for, respectively. It is shown how simple models of speciation and extinction lead directly to predictions of the structure of phylogenetic trees. These predictions include both essential attributes of phylogenetic trees: lengths, in the form of internode distances; and topology, in the form of internode links. These models also lead directly to statistical tests which can be used to compare predictions with phylogenetic trees that are estimated from data. Two different models and eight data sets are considered. A model without species extinction consistently yielded predictions closer to observations than did a model that included extinction. It is proposed that it may be useful to think of the diversification of recently formed monophyletic groups as a random speciation process without extinction.

Key words.—Coalescent, extinction, molecular clock, null model, speciation, species diversity.

Received March 6, 1991. Accepted October 23, 1991.

The sciences of ecology and evolutionary biology overlap in the study of the diversity of species. Researchers inquiring of the constraints on diversity can suppose that speciation and extinction are affected by a great variety of biotic and abiotic phenomena. Posing questions about species diversity in general would be impractical but for the tool of null models. The statistical comparison of observations with predictions from simple quantitative models can provide, in the case of a good fit, a practical approximation to reality; and, in the case of a poor fit, insight to particularly strong deterministic forces.

A common type of null model of taxonomic diversity assumes equanimity among lineages or taxonomic groups for speciation and extinction rates (Raup, 1985). Within this general class, models can take many forms. For example, simulations of random speciation and extinction have been used to predict clade shapes that are similar to those seen in the fossil record (Gould et al., 1977; Gilinsky and Good, 1989; Raup et al., 1973). In contrast, Dial and Marzluff (1989) addressed the observation of a hollow curve distribution of the number of taxonomic subunits (e.g., genera) within taxonomic units (e.g., families) that is found in a variety of extant vertebrates. They examined five null models and rejected all.

This report describes two very simple null models and shows how they lead to predictions about the structure of phylogenetic trees. These predictions concern both the lengths of trees and their topology, but in ways not typical of most phylogenetic analyses. Rather, these models have more in common with recent population genetic research of neutral models of gene genealogies (for reviews see Ewens, 1990; Hudson, 1990), in which all individuals are equally likely to pass on their genes. Similarly the models considered here can be considered as neutral speciation models, in which all species are equally likely to undergo speciation. Most important is that these models lead directly to simple goodness-of-fit tests which can be applied to empirical data. This report concludes with the results of goodness-of-fit tests on eight data sets.

Theory

Two Models of Diversification.—Consider the simple Markov process described by Yule (1924). In this model (model G for growth) there is no extinction, and for every point in time, all species are equally likely to undergo speciation. The time to speciation for each species follows an exponential probability distribution so that the rate of speciation, a, is the parameter of the exponential distribution. Put another way, the probability that the time until speciation is t is given by

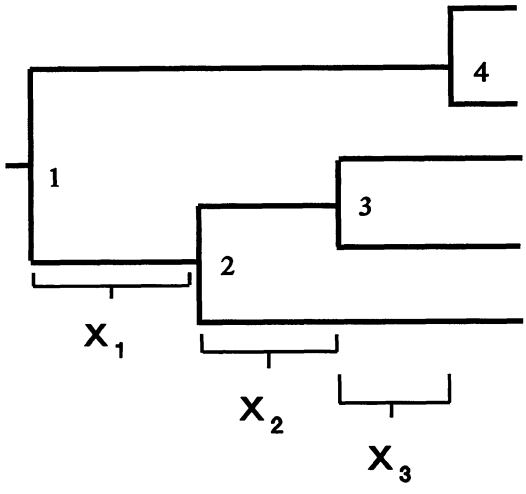

FIG. 1. A phylogenetic tree for five species. Node numbers and internode distances are indicated.

$$P(t) = ae^{-at}. \tag{1}$$

The exponential distribution has the Markov property whereby the distribution of the time of future events does not depend on the time of prior events. Thus the probability that the next speciation event occurs at time t is completely independent of the time since the last speciation event. It can also be shown that when there are N species, the overall speciation rate is Na and the probability distribution of the time until any one of the species undergoes speciation follows an exponential distribution with parameter Na.

Consider a monophyletic group of N extant species (i.e., N includes all of the extant species that have descended from an ancestral species). Figure 1 shows an example of a phylogenetic tree with both the nodes and the internode distances ordered in time. The nodes represent speciation events and the internode distances represent the time that has passed between successive speciation events. In general, a tree for N species will have $N - 1$ nodes and $N - 2$ internode distances. If nodes are ordered as in Figure 1, then $i + 1$ lineages extend between node i and node $i + 1$, where $1 \leq i \leq N - 1$. The internode distance between node i and node $i + 1$ will be referred to with the variable x_i (Fig. 1). Since the internode distances are equivalent to the times between successive speciation events, these may be compared with predictions from model G.

Under model G, each of the $N - 2$ internode distances is a random variable. The $N - 2$ exponential distributions all share the parameter a but differ in the multiplier of a. Thus, under the model, the time between node i and node $i + 1$ is described by an exponential distribution having parameter $(i + 1)a$.

To compare the predictions of model G with actual phylogenetic trees, a method is required to calculate a value for a that is consistent with observations. One approach is to generate a maximum likelihood estimate of a using internode distances from an actual phylogenetic tree. This approach requires a likelihood function, representing the overall likelihood of a set of observed internode distances, assuming that model G is correct. Since, under the model, each of the internode distances is an independent random variable, the likelihood function is the product of multiple exponential terms. The likelihood function for a set of internode distances from a tree of N species is

$$\begin{aligned} L(a) &= P(x_1, x_2, \ldots, x_{N-2}; a) \\ &= ae^{-ax_1} \cdot 2ae^{-2ax_2} \cdot \ldots \\ &\quad \cdot (N-2)a^{-(N-2)ax_{N-2}} \\ &= a^{N-2}(N-2)!e^{-a\sum_{i=1}^{N-2}(i+1)x_i}. \end{aligned} \tag{2}$$

The maximum likelihood estimate of the speciation rate, $\hat{a}$, follows directly:

$$\hat{a} = \frac{N-2}{\sum_{i=1}^{N-2}(i+1)x_i}. \tag{3}$$

Like most maximum likelihood estimators, expression (3) is biased. An unbiased estimate can be obtained by using

$$\hat{a}_u = \frac{N-3}{\sum_{i=1}^{N-2}(i+1)x_i}. \tag{4}$$

This quantity differs very little from expression (3) unless N is small.

Model G can be modified to include extinction by drawing on recent population genetic theory. Coalescent models (Kingman, 1982*a*, 1982*b*; Tajima, 1983; Tavare, 1984) in population genetics begin with a model of a population persisting via some

demographic process that proceeds forward in time, for example a Wright-Fisher model (Fisher, 1930; Wright, 1931), and then proceed to consider the history of a sample of genes from that population as a genealogical process that extends into the past. Similarly, we may consider a group of species that persist via some process of speciation and extinction, and then we may consider the history of a sample of extant species. The structure of the history depends on the model of diversification. As time moves forward, the relevant events are speciation and extinction. Looking backwards through time with a phylogenetic tree, the relevant events are nodes of common ancestry, indicating some subset of the speciation events.

Let the time until speciation for each of N species follow an exponential distribution with parameter B. Suppose that whenever any one of the species undergoes speciation, one of the others goes extinct. Thus N is constant over time and the time between any successive pair of speciation/extinction events is exponential with parameter NB. We will refer to this model as model C (for constant). Unlike model G, in which the nodes of the phylogenetic tree describe all speciation events, the nodes of the phylogenetic tree for a group of extant species under model C will correspond to only a subset of speciation/extinction events. Only those speciation/extinction events for which both lineages appear in the sample will be represented as nodes in the phylogenetic tree. For example, if one or both of the lineages that arise from a single lineage at the time of speciation subsequently go extinct, then that speciation event cannot be represented in the phylogenetic tree of extant species. In the case of model G, the time between speciation events corresponds to internode distance on the phylogenetic tree. In model C, because of extinction, the time between speciation/extinction events does not directly correspond to internode distance.

The distribution of internode distances can be obtained from the distribution of times between speciation/extinction events. Consider the history of a monophyletic group of N extant species, and let the most recent speciation/extinction event be denoted as event one, the second most recent event as event two and so on into the past. For the moment permit the assumption that the time at which species are sampled (i.e., the present) coincides with a speciation/extinction event. This assumption ensures that the time between the present and event one also follows an exponential distribution with parameter NB. Then the probability distribution of the time between the present and event j, T_j, is the distribution of the sum of j identically distributed exponential variables. The distribution of the sum can be found via convolution of the component distributions, which in this case yields the gamma distribution

$$P(T_j = t_j) = NBe^{-NBt_j}\frac{(NBt_j)^{j-1}}{(j-1)!}. \quad (5)$$

Now consider a phylogenetic tree for a random sample of n extant species that are taken from the group of N species, where $n \leq N$ (model C is more general than model G, wherein the latter requires that the sample of species be an entire monophyletic group while the former does not). As in model G, the tree has $n - 1$ nodes indexed so that $i + 1$ lineages extend between node i and node $i + 1$. Note that the index of speciation/extinction events, j, increases for events further in the past while the index of nodes, i, decreases for events further in the past. Let M_i represent the speciation/extinction event that is associated with node i. Let $P(M_i = j)$ be the probability that node i corresponds to speciation/extinction event j. If $n = N$, then $P(M_{n-1} = 1) = 1$. This is because the two species from the most recent speciation/extinction event must be included in the sample when $N = n$. For $n < N$, calculation of $P(M_{n-1} = 1)$ follows from the consideration that in order for node $n - 1$ to coincide with the most recent speciation/extinction event, the pair of species created at that speciation/extinction event must be included in the random sample of n species. The probability of this is the hypergeometric probability

$$\frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}. \quad (6)$$

More generally, the probablity, γ_i, that $i +$

1 species include a pair that had a most recent common ancestor at the most recent node (i.e., node i) can be represented by

$$\gamma_i = \frac{\binom{N-2}{(i+1)-2}}{\binom{N}{i+1}} = \frac{i(i+1)}{N(N-1)}. \quad (7)$$

It follows that $P(M_{n-1} = 2) = (1 - \gamma_{n-1})\gamma_{n-1}$ and, in general, the relationship between speciation/extinction events and phylogeny nodes is described by a geometric distribution:

$$P(M_{n-1} = j) = (1 - \gamma_{n-1})^{j-1}\,\gamma_{n-1}. \quad (8)$$

Let $P(x_{n-1})$ be the probability that the interval between T_0 and node $n - 1$ has length x_{n-1}. Then

$$\begin{aligned} P(x_{n-1}) &= \sum_{j=1}^{\infty} P(M_{n-1} = j)P(T_j = x_{n-1}) \\ &= \sum_{j=1}^{\infty} (1 - \gamma_{n-1})^{j-1}\gamma_{n-1}NB \\ &\qquad \times (NBx_{n-1})^{j-1}\frac{e^{-NBx_{n-1}}}{(j-1)!} \\ &= NB\gamma_{n-1}e^{-NB\gamma_{n-1}x_{n-1}}. \end{aligned} \quad (9)$$

Evidently the probability density is an exponential distribution with parameter $NB\gamma_{n-1}$, and thus the process of proceeding backwards in time to successive nodes is Markovian. The reasoning that led to expression (9) can be repeated beginning at node $n - 1$ instead of T_0. Regardless of the time between T_0 and node $n - 1$, the time between node $n - 1$ and node $n - 2$ will follow an exponential distribution having parameter $NB\gamma_{n-2}$. In general, we can treat the time interval between nodes i and $i + 1$ as if beginning with a random sample of $i + 1$ species. Actually the time between T_0 and node $n - 1$ should not be used unless T_0 coincides with a speciation/extinction event. However, for node i, where $1 \leq i \leq n - 2$, the probability density is exponential with parameter $NB\gamma_i$.

If we let $b = B/(N - 1)$ and $\lambda_i = i(i + 1)$ then $NB\gamma_i = b\lambda_i$. Thus the parameter for the probability density of x_i is seen to consist of two components; b is a parameter of the species group and λ_i is determined by the sample size.

As with $\hat{a}$ for model G, the maximum likelihood estimate of b can be calculated from a set of observed internode distances:

$$\hat{b} = \frac{n-2}{\sum_{i=1}^{n-2} \lambda_i x_i}. \quad (10)$$

Expression (10) is nearly identical to expression (2). Also, as for $\hat{a}$, an unbiased estimate of b can be obtained using

$$\hat{b}_u = \frac{n-3}{\sum_{i=1}^{n-2} \lambda_i x_i}. \quad (11)$$

Although model C includes extinction and model G does not, the two models fit a common mathematical framework. In both models, the steps of a simple Markov chain can be used to model the times between nodes of a phylogenetic tree. For different reasons, neither model requires estimating the total number of species. In model G, all species are included in the sample, so that the number of species is known. In model C, the total number of species need not be known because it is a component of the parameter b. This has the advantage of permitting the use of a random sample of species, but has the disadvantage that unless the total number of species is known, the speciation rate, B, cannot be estimated.

For identical speciation rates and the same number of species the two models predict similar distances for recent nodes. In Figure 2 the expected internode distances are plotted under both models for $N = 20$ and $a = B = 1$. At earlier times the predictions of the two models become increasingly different. In general, the discrepancy between models with extinction versus those without will be greater for events further in the past. The sample is necessarily restricted to lineages that did not go extinct, and as the number of sampled lineages is increasingly small further in the past, the internode distances are lengthened by events involving lineages not included in the sample.

Model C differs from a more general model in which speciation and extinction

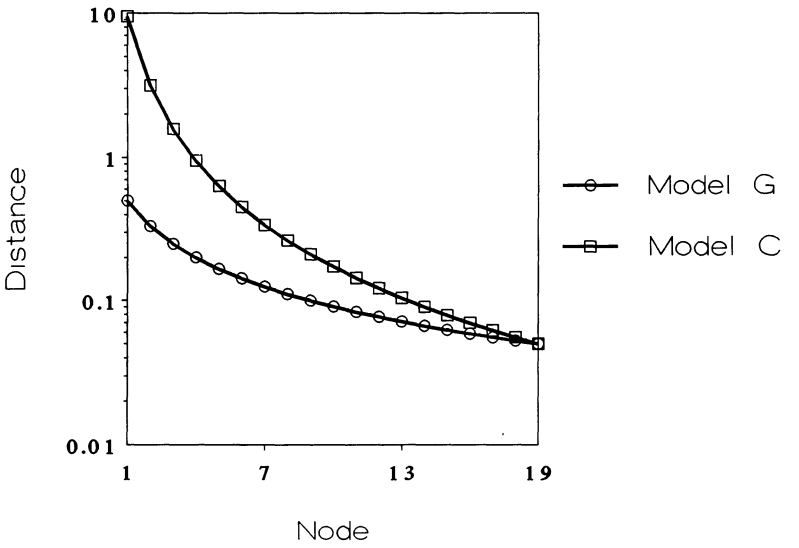

FIG. 2. The natural logarithms of expected internode distances for models G and C are compared. Each case begins with a sample of 20 species. The speciation rate per species was set at 1 for both models (i.e., $a = B = 1$). Because the expected value of an exponential distribution is always the inverse of the parameter of that distribution, the expected distance between nodes i and $i + 1$ for model G is equal to $1/(i + 1)$. Similarly, for model C the expected distance between nodes i and $i + 1$ is equal to $1/(\lambda_i/19)$.

are independent random events (Raup, 1985). In the more general model the times between events (whether speciation or extinction) can be modeled as a birth and death Markov chain. However, this model does not lead to predictions of internode distances as simply as does model C. It is anticipated that were speciation and extinction modeled as independent events, the expectation for internode distances would be similar to that for model C in Figure 2. This is because the same reasoning on increasing internode distances because of extinction applies to both models. However, the actual distribution of internode distances under the more general model may be very different from model C.

Models G and C are readily compared with simulations. To simulate a data set, all that is required is a set of random numbers drawn from exponential distributions. For model G a simulated value for the distance between node i and node $i + 1$ requires a random number drawn from an exponential distribution having parameter $(i + 1)a$. The distance for the same node under model C is created by drawing from an exponential distribution having parameter $b\lambda_i$. Figures 3A and 3B show the results of single simulations of internode distances, under models G and C, respectively. For both sets of distances $\hat{a}$ and $\hat{b}$ were calculated, and these were used to plot the expected values assuming models G and C, respectively. Knowing that model C predicts a steeper curve (Fig. 2) we expect the lines of best fit to cross, when both models are applied to a data set (simulated or real). In Figure 3 the lines of best fit cross as expected, and model G appears to fit best in Figure 3A while model C appears to fit best in Figure 3B.

The maximum likelihood approach to parameter estimation, used in both models, leads readily to goodness-of-fit tests. The goodness-of-fit of predictions to observations can be described with a likelihood ratio statistic. Under model G, for a particular internode distance, x_i, the maximum likelihood estimate of a, $\hat{a}_i$, is $1/(ix_i)$. The likelihood ratio statistic for model G is the ratio of the likelihood using $\hat{a}$ to the likelihood using $\hat{a}_i$:

$$L_G = \frac{\prod_{i=1}^{N-2} \hat{a} i e^{-\hat{a} i x_i}}{\prod_{i=1}^{N-2} \hat{a}_i i e^{-\hat{a}_i i x_i}}. \qquad (12)$$

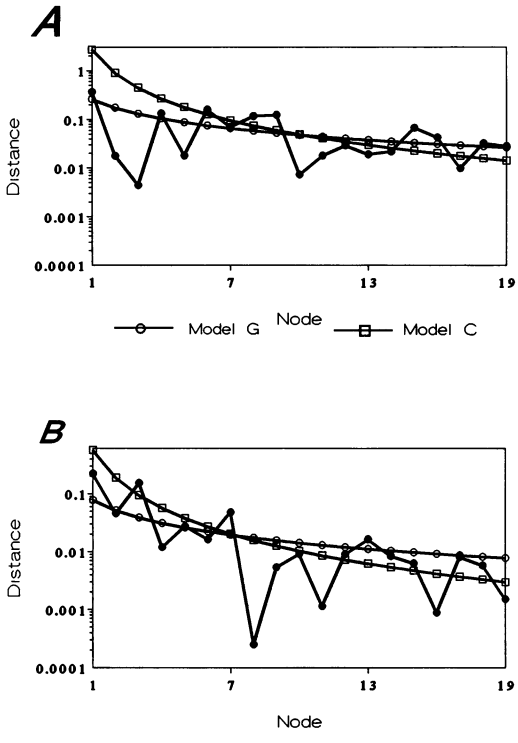

FIG. 3. The natural logarithms of simulated internode distances for phylogenetic trees of 20 species. Simulated distances are plotted with the best fit values assuming models G and C. A shows the results of a model G simulation. The value for the distance between nodes i and $i + 1$ was randomly drawn from an exponential distribution with parameter $(i + 1)$ (i.e., $a = 1$). The maximum likelihood estimate of the parameter assuming models G and C were $\hat{a} = 1.91$ ($\Lambda_G = 14.0$) and $\hat{b} = 0.19$ ($\Lambda_C = 22.4$), respectively. B shows the results of a model C simulation. The value for the distance between nodes i and $i + 1$ was randomly drawn from an exponential distribution with parameter λ_i (i.e., $b = 1$). The maximum likelihood estimate of the parameter assuming models G and C were $\hat{a} = 6.41$ ($\Lambda_G = 22.4$) and $\hat{b} = 0.88$ ($\Lambda_C = 19.2$), respectively. For both A and B, the best fit values of the distance between nodes i and $i + 1$ were equal to $1/[\hat{a}(i + 1)]$, under model G, and equal to $1/(\hat{b}\lambda_i)$, under model C.

In practice, it is convenient to use

$$\Lambda_G = -2 \ln(L_G)$$

$$= -2\left[(n - 2) \ln(\hat{a}) + \sum_{i=1}^{n-2} \ln(ix_i)\right]. \quad (13)$$

A direct way to test whether or not an observed value for Λ_G is greater than expected under a neutral model is to calculate $\hat{a}$ from the data, and to compare the corresponding value of Λ_G with the results of simulations. Simulated data for a set of N species can be created by drawing a single random number from each of $N - 2$ exponential distributions. The distance for node i is drawn from an exponential distribution having parameter $\hat{a}(i + 1)$. For each simulation (i.e., each set of $N - 2$ random numbers) $\hat{a}$ and Λ_G are calculated. After many simulations, the values of Λ_G are ranked, and the location of the observed value of Λ_G within the simulated distribution is found. The values of $\hat{a}$ generated by the simulations can be used to generate a variance and 95% confidence limits for the observed $\hat{a}$. In the course of simulations, it was found that $\Lambda_G/1.17$ has a distribution nearly identical to the χ^2 distribution with $N - 3$ degrees of freedom. This means that the χ^2_{N-3} distribution can be used for significance testing.

For model C, the analogous expression for Λ_C is identical to expression (13) except that $\hat{b}$ replaces $\hat{a}$ and λ_i replaces i. The procedure for testing the fit of model C follows that for model G exactly, including the use of the χ^2_{N-3} distribution.

For the simulations in Figure 3, values of Λ_G and Λ_C are given. The relative values of the likelihood ratios fit the expectation that the fit of model G is better than the fit of model C to a simulation of model G. Similarly model C fits the model C simulation better. The simulation in Figure 3 was picked for this reason. This pattern is expected for the majority, but not all, simulations. Sometimes a model G simulation will look more like an example of model C, and vice versa.

Topology Theory.—If all species in a monophyletic group are equally likely to undergo speciation, it is appropriate to view the particular species that does undergo speciation as a random selection from the group. One way to examine departures from randomness is to consider that at any point in time the species of a monophyletic group may be divided into two groups: those two species that arose from the last speciation event; and all others. Under both models G and C, the two species that arise from one speciation event should be no more or less likely to undergo the next speciation event than are other species in the group. Simply by examining a phylogenetic tree, one can

determine, for each node, whether the lineage that underwent speciation had arisen from the previous node. Since one of the two species that arise from node 1 must be involved in node 2, only the $N - 3$ internodes between nodes 2 and N $-$ 1 can be included in a test for departure from randomness.

The probability that one of the two species arising from the speciation event at node i connects to node $i + 1$ is $2/(i + 1)$. Note that this quantity decreases for increasing node numbers. The probability that one of these two species does not connect is $1 - 2/(i + 1) = (i - 1)/(i + 1)$. Thus, except for links from nodes 2 to 3 and from nodes 3 to 4, a new species to new species link is always less likely than the alternative.

We may describe the pattern between node 2 and node $N - 1$ of a phylogenetic tree (hereafter referred to as a Links pattern) as a series of 1's and 0's, with 1's representing cases where successive speciation events are linked. For example, consider the tree of five species in Figure 1. For nodes 2 and 3 the Links patterns is 1·0, meaning that for node 2, but not node 3, one of the new species was connected to the next node. Since, under models G and C, the linkages between successive nodes are independent of each other, the likelihood of an entire Links pattern (hereafter referred to as a Links likelihood) is the product of the likelihoods of the observations at individual nodes. For the tree in Figure 1 the likelihood is $(2/3)\cdot(2/4) = 1/3$. To formalize this calculation let l_j be the value in the Links pattern at position j (i.e., corresponding to node $j + 1$ in the phylogenetic tree). Then let

$$f(l_j) = \frac{2}{j + 2}, \qquad \text{for } l_j = 1, \quad (14)$$

and

$$f(l_j) = \frac{j}{j + 2}, \qquad \text{for } l_j = 0. \quad (15)$$

The Links likelihood can be expressed as

$$\prod_{j=1}^{N-3} f(l_j). \quad (16)$$

There are two patterns of nonrandomness to which a Links test should be sensitive: the case where new species are more likely than others to undergo speciation; and the case where new species are less likely to undergo speciation. For most nodes a "1" in a Links pattern has a likelihood less than or equal to 0.5 while a "0" has a likelihood greater than or equal to 0.5 (the exception is node 2). Thus a tree with many 1's will have a low likelihood and a tree with many 0's will have a high likelihood. The Links likelihood can be used as a test statistic because both low and high values correspond to circumstances to which a test should be sensitive. The Links likelihood is imperfect because of node 2, at which a new species to new species link has a likelihood of 2/3. This means a Links pattern with a "1" for node 2 will have a higher likelihood than the same pattern with a "0" for node 2. Thus the Links likelihood is a better test statistic when trees are large and node 2 contributes relatively little.

A test of the departure of a Links pattern from randomness is made by contrasting the observed value for the Links likelihood with the distribution of possible values. Because each of the $N - 3$ links could be a "1" or a "0," there are $2^{(N-3)}$ distinct possible Links patterns. A test is made by calculating a Links likelihood for each possible pattern, and summing those that are equal to or less than that for the actual observation. A summed likelihood of $P \leq 0.05$ (0.025 in two tailed test) would indicate that the probability of getting an equally or more extreme pattern was only 0.05, suggesting new species undergo speciation more than is expected. A summed likelihood of $P \geq 0.95$ (0.975 in a two tailed test) would suggest that new species undergo speciation less than is expected. When N is large, P can be calculated from a random sample of the possible observations. The Links test is more sensitive to excessive new species to new species links. A comblike tree (i.e., an observation of all 1's) is not statistically significant unless there are nine or more species (i.e., six or more 1's), while an observation of all 0's is not statistically significant unless there are twenty-one or more species.

The Links test is just one of many ways in which topologies could be examined for randomness. For example, Slowinski and

Gruyer (1989) describe a test for imbalance between the sizes of sister groups that extend from the basal node of a tree. This test is intended for cases where one or both sister groups is large.

Testing Models

If the estimate of a phylogenetic tree is not biased with respect to the models, then we can use estimated trees for goodness-of-fit tests. However, even if they are unbiased, estimated trees will have an extra source of variance (i.e., the departure of estimated trees from true trees) not accounted for by models G and C. This means that we can expect to find a poor fit of estimated trees, even when true trees may be consistent with one of the models, more often than indicated by the chosen significance level.

As they are described in the section on *Theory,* models G and C can be used to generate predictions of the times between nodes. However, most estimated phylogenies contain branch lengths in units of change of the characters used to construct the tree. These measures of branch lengths can be converted to internode distances only if they are directly proportional to time. The effect, of using distances that are proportional to time rather than equal to time, is that the x_i values are in units of character change rather than in units of time.

The remainder of the analysis employs the assumption that for the characters used in the construction of a phylogenetic tree, species divergence proceeds without homoplasy and at a constant rate (i.e., in a clocklike fashion). If this assumption is not made, then application of models G and C to actual data becomes much more difficult. In essence, without the clock assumption, an estimated phylogenetic tree cannot inform on the time between nodes. While there may be circumstances in which data on variation in evolutionary rates can be used in conjunction with estimated phylogenies, this approach has not been pursued here. Rather, the clock assumption is made, and analysis is restricted to those data sets that seem most consistent with it (see below).

Studies that compared methods of phylogeny estimation under the assumption of clocklike divergence, and across trees of wide ranging topologies and branch lengths, have found that the UPGMA (Sneath and Sokal, 1973) clustering method generally performed as good as and often better than parsimony and maximum-likelihood methods (Rohlf and Wooten, 1988; Rohlf et al., 1990). The UPGMA method was selected for this reason as well as for ease of calculation and applicability to a wide variety of data sets. The NTSYS (Rohlf, 1985) package of programs was used for generating UPGMA trees and cophenetic correlations.

Each of the data sets examined satisfied three criteria: the species formed a monophyletic group, in the sense that all extant species of a clade are included; the distance data were consistent with clocklike divergence; and sufficient data were presented to distinguish all species of the group (i.e., data sets were excluded if zero divergence was reported between some species). A monophyletic group is required for testing model G. Model C requires either a monophyletic group or a random sample from a monophyletic group. Tests were restricted to molecular (i.e., RFLP, DNA-DNA hybridization, and protein electrophoresis). These types of data are often consistent with a model of a constant rate of change (Caccone and Powell, 1987; Caccone et al., 1988; Easteal, 1988*a*, 1990).

Molecular data may underestimate branch lengths for early nodes if nucleotide sites become saturated with mutations over time. For RFLP data and DNA-DNA hybridization data, this bias is expected to be slight for the low levels of divergence reported in the data sets used here. For allelic (e.g., protein electrophoresis) data, however, saturation can be a major concern. With two exceptions, this type of data was not used. The *Ursus* (Goldman et al., 1989) data includes both allozyme and two-dimensional electrophoresis data and genetic distances are low. The *Plethodon* (Highton and Larson, 1979) data set has some high genetic distances, but the authors provide additional data showing that their estimates of genetic distance, based on allozyme comparisons, are linearly correlated with DNA-DNA hybridization data, even for high levels of genetic distance.

An extensive literature search yielded exactly eight studies that satisfied the criteria. In each of these cases monophyly was as-

sumed if all extant species of a taxon are included in the study. It cannot be ruled out that in some cases not all extant species are recognized or that a taxon is actually paraphyletic. In the case of sect. *Peripetasma* of the genus *Clarkia* a member of a different monotypic genus, *Heteroguara,* that had been included in the study, actually clustered within *Peripetasma* (Sytsma and Gottlieb, 1986). Thus sect. *Peripetasma* appears to be paraphyletic. For this analysis *Heteroguara* was included within *Peripetasma.*

If divergence were strictly clocklike, then a matrix of all pairwise distances between N species would contain only $N - 1$ distinct values, corresponding to the times of the $N - 1$ nodes. In this case the cophenetic correlation, the correlation between the actual pairwise distances and those generated using the distances in the UPGMA tree, would be equal to one. The cophenetic correlations for the data sets are listed in Table 1. All values are above 0.877. Interestingly, the values for DNA-DNA hybridization studies are all higher than those for the RFLP studies which are in turn higher than those for the protein studies. Some authors addressed the issue of an evolutionary clock in their data (sect. *Peripetasma*: Sytsma and Gottlieb, 1986; tribe Hadenoecini: Caccone and Powell, 1987; subgroup *melanogaster*: Caccone et al., 1988*a*; genus *Ursus*: Goldman et al., 1989; family Gruidae: Krajewski, 1990). With the exception of a departure from rate constancy in one lineage of the Gruidae (Krajewski, 1990), clocklike divergence could not be rejected in these studies.

Table 1 reveals that in no case can either model G or model C be rejected on the basis of goodness-of-fit to the data. In all eight cases, model C fits worse (i.e., has a higher value for Λ) than does model G. The consistently poorer fit of model C suggests that if the results of the eight tests are taken together model C may be rejected. However, the result of Fisher's (1954) test of combined probabilities yields a test statistic of 20.61, which is not significant (from χ^2_{16}, $P = 0.194$). Thus neither model can be rejected outright. The actual and predicted internode distances are plotted in Figure 4.

Despite the results from the goodness-of-fit tests, there remains a sense in which the consistently worse fit of model C is unlikely. Consider the null hypothesis that for any particular data set there is a fifty-fifty chance that model C fits better than model G. Under this hypothesis the probability of the outcome, in which for all eight cases one model fit worse than the other, is small ($0.5^8 = 0.004$). Thus taken together, the eight data sets show that model C yields a poorer fit more often than does model G. The Links tests reveal no significant deviations from random expectations (Table 2). In light of the finding that model G consistently fits better than does model C, the Links test is expected to be more useful. Any underlying pattern of increased or decreased likelihood on the part of new species to undergo speciation will be obscured to the extent that extinction of some species occurs.

Speciation Rates

The units of measure of the RFLP and DNA-DNA hybridization data sets are comparable. In the RFLP studies, distance was expressed as the percentage of sequence divergence (Nei and Li, 1979). The hybridization studies all use ΔT_m for which a value of 1° has been equated with 1.7% sequence divergence (Caccone et al., 1988*b*). The average of $\hat{a}$ for these six data sets (with those values for the DNA-DNA hybridization studies adjusted by 1.7) is 0.66. In other words the average estimated rate of speciation is 0.66 speciation events per 1% sequence divergence. Put another way, the inverse indicates that approximately 1.5% sequence divergence occurs between speciation events. If substitutions accumulate at $5 \cdot 10^{-9}$ per site per year, then these numbers correspond to a speciation rate of one event per three million years, on average. These values should not be considered as estimates of global parameters, since the wide ranging and sometimes nonoverlapping 95% confidence limits indicate significant variation for a.

The same analysis applied to model C reveals an average value of $\hat{b}$ of 0.14. Recall that $b = B/(N - 1)$. Since this study was limited to complete monophyletic groups, the values for N can be used to convert estimates of b to estimates of B. These calculations reveal an average speciation rate of 1.23 events per 1% divergence. Thus

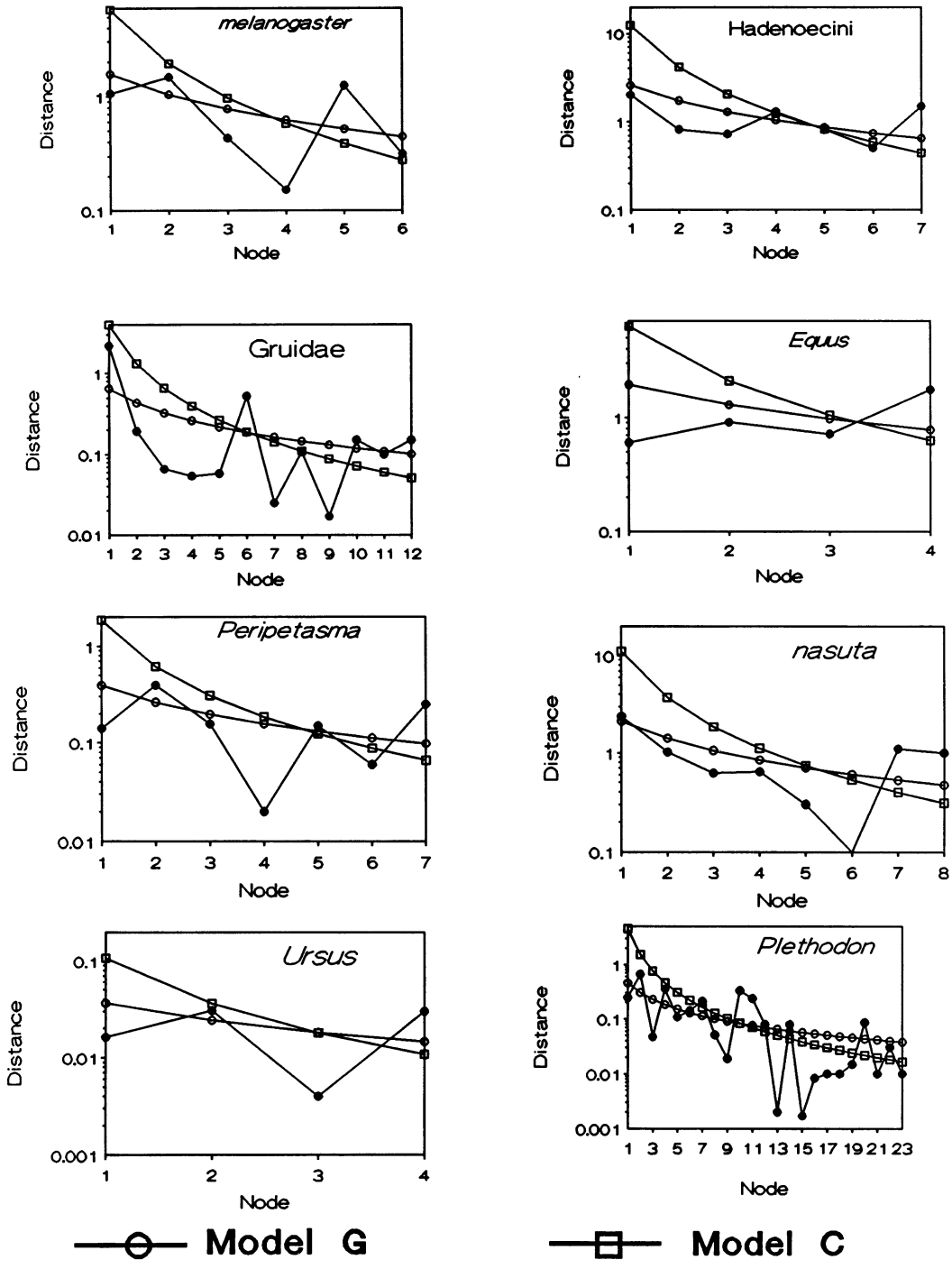

FIG. 4. The natural logarithms of internode distances obtained from UPGMA trees of the data sets listed in Table 1. The expected values for node i under models G and C are equal to $1/[\hat{a}(i+1)]$ and $1/(\hat{b}\lambda_i)$, respectively.

TABLE 1. Neutral models fit to phylogenetic trees. N is the number of species in the tree. The number of internode distances used for models G and C is equal to $N - 2$. P is the probability of getting a likelihood ratio statistic equal to or greater than that observed under the model. Values of $\hat{a}$ and $\hat{b}$ were calculated using expressions (3) and (10), respectively. The internode distances for these calculations were generated by applying the UPGMA protocol in the SAHN program of NTSYS (Rohlf, 1985) to the distance data reported in the references. The cophenetic correlations, r, were calculated using the programs SAHN, COPH, and MXCOMP in NTSYS. The 95% confidence limits of $\hat{a}$ were calculated as follows: using the observed value of $\hat{a}$, 7,500 simulated data sets were created; for each simulated data set, $\hat{a}$ was calculated; the simulated values were ranked; and the upper and lower limits were taken from the 97.5% and 2.5% positions in the ranking, respectively. The same procedure was done for the 95% confidence limits of $\hat{b}$. Subgroup *melanogaster* (Caccone et al., 1988*a*): ΔT_m data for all species of these members of the genus *Drosophila*. Hadenoecini (Caccone and Powell, 1987): ΔT_m data from all species of this tribe of the family Raphidophoridae. Gruidae (Krajewski, 1990): ΔT_m data for all 14 species of this family of class Aves. Data for the four subspecies of *D. sulfurigaster* were pooled because of zero divergence among them. *Equus* (George and Ryder, 1986): sequence divergence based on restriction site analysis of mitochondria DNA of family Equidae. Section *Peripetasma* (Systma and Gottlieb, 1986): sequence divergence based on restriction site analysis of chloroplast DNA for all nine species of this section of the genus *Clarkia* (Onagraceae). Subgroup *nasuta* (Chang et al., 1989): sequence divergence based on restriction site analysis of mitochondria DNA for these members of genus *Drosophila*. *Ursus* (Goldman et al., 1989): genetic distance measures (Nei, 1978) from two-dimensional protein electrophoresis and allozyme data where summed for six species in the family Ursidae. Data for two other genera of this family were not used because of fossil evidence of speciation and extinction in these lineages (Goldman et al., 1989). *Plethodon* (Highton and Larson, 1979): genetic distance (Nei, 1978) from allozyme data for all members of this genus from the family Plethodontidae. Two nodes had identical branch lengths of 0.32. Because an internode length of zero can not be included in the calculation of the goodness of fit, this internode distance was not used. Thus all numbers for both G and C models are actually based on 23 internode distances.

				Model G				Model C			
Data sets	Data type	N	r	$\hat{a}$	95% limits	Λ	P	$\hat{b}$	95% limits	Λ	P
subgroup *melanogaster*	DNA-DNA	8	0.990	0.32	0.17–0.87	3.01	0.77	0.09	0.05–0.24	5.65	0.44
tribe Hadenoecini	DNA-DNA	9	0.993	0.20	0.11–0.50	1.81	0.96	0.04	0.02–0.10	6.58	0.47
family Gruidae	DNA-DNA	14	0.981	0.77	0.47–1.48	13.44	0.40	0.13	0.08–0.26	16.73	0.22
genus *Equus*	RFLP	6	0.928	0.26	0.12–0.96	2.03	0.63	0.08	0.04–0.30	5.06	0.24
sect. *Peripetasma*	RFLP	9	0.950	1.27	0.68–3.18	4.96	0.65	0.27	0.14–0.66	9.59	0.23
subgroup *nasuta*	RFLP	10	0.917	0.24	0.13–0.54	4.34	0.81	0.04	0.02–0.11	9.71	0.31
genus *Ursus*	Protein	6	0.877	13.79	6.3–49.6	2.70	0.52	4.63	2.11–16.9	5.08	0.24
genus *Plethodon*	Protein	26	0.886	1.11	0.77–1.76	30.35	0.24	0.11	0.08–0.18	33.56	0.15

TABLE 2. Links tests.

Data sets	Links pattern	Likelihood	P^*
subgroup *melanogaster*	10001	0.0381	0.49
tribe Hadenoecini	000001	0.0119	0.26
family Gruidae	11000010001	0.0013	0.36
genus *Equus*	111	0.1333	—
sect. *Peripetasma*	100100	0.0357	0.76
subgroup *nasuta*	1111010	0.0062	0.22
genus *Ursus*	110	0.2000	—
genus *Plethodon*	000010000010 00000000100	$2.308 \cdot 10^{-5}$	0.40

* P is the probability of getting an observation with a likelihood equal to or less than the observation. This probability was not calculated for the smallest data sets, for which the Links likelihood is not a good test statistic (see text).

model C indicates an average speciation rate about twice that of model G.

DISCUSSION

Estimated phylogenetic trees are generally used for inferring historical patterns of common ancestry and to a lesser extent estimating the time of speciation events. It is shown here how estimated phylogenetic trees can be used to study processes of diversification. As is common with historical inference, information comes in the form of contrasts between observations and null predictions.

The consistently good fit of model G and the apparent randomness of the Links patterns suggest that it is useful to think of the diversification of species as a random process of speciation without extinction. In the face of evidence of extinction, this model is clearly untenable for many groups of organisms. The discrepancy, that model G fits well in this analysis but cannot be generally true, may be reconciled if we consider that the eight monophyletic groups studied are not typical. Two criteria of this analysis favored small recently formed monophyletic groups. First, the analysis was limited to studies of complete monophyletic groups and more such studies exist for small groups than for large ones. Second, the characters used by the different studies (protein electrophoretic mobility and DNA sequences) evolve quickly and are less applicable for older groups. With the exception of the study of sect. *Peripetasma* (Sytsma and Gottlieb, 1986) all of the studies calculated or reported rough estimates of the age of their respective groups. For the genus *Plethodon* this time was 40 million years (Highton and Larson, 1979); for the family Gruidae (Krajewski, 1990) the time was 23 million years; and the others were all less than 10 million years. In sum, a model that excludes extinction cannot be generally true, but it may be true of small groups of recent origin. If this is true, it must follow that extinction is a more likely event (per unit time) for older lineages.

Models G and C are minimal in their assumptions. This quality is desirable in null models for two reasons. The first is that predictions are more easily grasped in a cognitive sense. Thus, in Model G, the longer internode distances that arise when there are fewer species can be understood as a waiting time problem (i.e., waiting for any one of few events takes longer on average than does waiting for any one of many events). The second reason is that when simple models fail it is often possible to pinpoint which of the properties of the model are inaccurate.

Models G and C lead to simple statistical tests. Expressions (3), (10), and (13) are easily calculated, and the Links test requires a short computer program (computer programs for the Links test and the goodness-of-fit simulations can be obtained by a request to the author). The power of the statistical tests is necessarily low for small samples, and, because all species of a taxon are required, this is an important limitation of these tests for some groups of organisms. As the number of data sets involving entire monophyletic groups increases, these tests will become more useful.

It must be stressed that models G and C

provide predictions about time. These models can be applied to data in units of change only if change is directly proportional to time. It follows that these models should not be applied to data not consistent with clocklike divergence. In the case of clocklike divergence, it is simply a fortunate convenience that the scaler of proportionality need not be known and does not enter into the statistical tests. Although UPGMA was used here, any of a variety of phylogenetic estimation procedures could be used, so long as they return values interpretable as internode distances proportional to time. By far the most appropriate data would be from monophyletic groups with fairly accurate estimates of the actual times of divergence.

Acknowledgments

Thanks to three anonymous reviewers for many constructive comments. Thanks to P. Smouse for his help in developing the goodness-of-fit tests. This research was supported in part by NSF grant BSR-8918164.

Literature Cited

Caccone, A., G. D. Amato, and J. R. Powell. 1988*a*. Rates and patterns of scnDNA and mrDNA divergence within the *Drosophila melanogaster* subgroup. Genetics 118:671–683.

Caccone, A., R. DeSalle, and J. R. Powell. 1988*b*. Calibration of the change in thermal stability of DNA duplexes and degree of base mismatch. J. Mol. Evol. 27:212–216.

Caccone, A., and J. R. Powell. 1987. Molecular evolutionary divergence among North American cave crickets. II DNA-DNA hybridization. Evolution 41:1215–1238.

Chang, H., D. Wang, and F. J. Ayala. 1989. Mitochondrial DNA variation in the *Drosophila nasuta* subgroup of species. J. Mol. Evol. 28:337–348.

Dial, K. P., and J. M. Marzluff. 1989. Nonrandom diversification within taxonomic assemblages. Syst. Zool. 38:26–37.

Easteal, S. 1988. Rate constancy of globin gene evolution in placental mammals. Proc. Nat. Acad. Sci. 85:7622–7626.

———. 1990. The pattern of mammalian evolution and the relative rate of molecular evolution. Genetics 124:165–173.

Ewens, W. J. 1990. Population genetics theory—The past and the future, pp. 177–227. *In* S. Lessard (ed.), Mathematical and Statistical Developments of Evolutionary Theory. Kluwer Academic Publishers, Dordrecht, Germany.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection. Oxford University Press, London, UK.

———. 1954. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, UK.

George, M., Jr., and O. A. Ryder. 1986. Mitochondrial DNA variation in the genus *Equus*. Mol. Biol. Evol. 3:535–546.

Gilinsky, N. L., and I. J. Good. 1989. Analysis of clade shape using queuing theory and the fast Fourier transform. Paleobiology 15:321–333.

Goldman, D., P. R. Giri, and S. J. O'Brien. 1989. Molecular genetic distance estimates among the Ursidae as indicated by one- and two-dimensional protein electrophoresis. Evolution 43:282–295.

Gould, S. J., D. M. Raup, J. J. Sepkoski, Jr., T. J. M. Schopf, and D. S. Simberloff. 1977. The shape of evolution: A comparison of real and random clades. Paleobiology 3:23–40.

Highton, R., and A. Larson. 1979. The genetic relationships of the salamanders of the genus *Plethodon*. Syst. Zool. 28:579–599.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, pp. 1–44. *In* P. H. Harvey and L. Partridge (eds.), Oxford Surveys in Evolutionary Biology, Vol. 7. Oxford University Press, N.Y., USA.

Kingman, J. F. C. 1982*a*. The coalescent. Stochastic Processes Appl. 13:235–248.

———. 1982*b*. On the genealogy of large populations. J. Appl. Probab. 19A:27–43.

Krajewski, C. 1990. Relative rates of single-copy DNA evolution in cranes. Mol. Biol. Evol. 7:65–73.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590.

Nei, M., and W.-H. Li. 1979. Mathematical models for studying genetic variation in terms of restriction endonucleases. Proc. Nat. Acad. Sci. 76:5269–5273.

Raup, D. M. 1985. Mathematical models of cladogenesis. Paleobiology 11:42–52.

Raup, D. M., S. J. Gould, T. J. M. Schopf, and D. S. Simberloff. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81: 525–542.

Rohlf, F. J. 1985. NTSYS. Numerical Taxonomy System of Multivariate Statistical Programs. State Univ. New York, Stony Brook, USA.

Rohlf, F. J., and M. C. Wooten. 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. Evolution 42:581–595.

Rohlf, F. J., W. S. Chang, R. R. Sokal, and J. Kim. 1990. Accuracy of estimated phylogenies: Effects of tree topology and evolutionary model. Evolution 44:1671–1684.

Slowinski, J. B., and C. Gruyer. 1989. Testing the stochasticity of patterns of organismal diversity: An improved null model. Am. Nat. 134:907–921.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical Taxonomy. Freeman, San Francisco, CA USA.

Sytsma, K. J., and L. D. Gottlieb. 1986. Chloroplast DNA evolution and the phylogenetic relationships in *Clarkia* sect. *Peripetasma*. Evolution 40:1248–1262.

Tajima, F. 1983. Evolutionary relationship of DNA

sequences in finite populations. Genetics 105:437–460.

TAVARE, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

WRIGHT, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

YULE, G. U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F. R. S. Philos. Trans. R. Soc. Lond. B:213:21–87.

Corresponding Editor: M. Donoghue