5

Testing Models of Speciation and Extinction with Phylogenetic Trees of Extant Taxa

Jody Hey, Holly Hilton, Nicholas Leahy,
and Rong-Lin Wang

Could the diversity of species have been caused by simple random processes of speciation and extinction? This question, asked repeatedly in recent decades (Raup et al. 1973; Gould et al. 1977; Stanley et al. 1981; Raup 1985; Hey 1992; Nee et al. 1992) and in this chapter, might seem irrelevant to investigators of specific cases of speciation or extinction. Clearly, every case of speciation or extinction has ecological, biogeographic, and genetic causes, and the idea of "random causes" may have little meaning to students of the mechanisms of speciation and extinction. However, scientists have few tools to study the causes of diversity for the vast majority of speciations and extinctions that happened long ago. The kinds of data that can be brought to bear on the causes of ancient events are mostly limited to just one type: assessments of the similarities of and differences between organisms. If the comparisons are among fossils of organisms, then some information on the age of fossils is also available. In general, these data offer no information on individual ancient speciation and extinction events, but they do present limited scope for inference on general causes of diversity. For instance, comparative phenotypic data (or genotypic data if species are extant) can be used to estimate a phylogenetic tree. We may then ask what kinds of processes could have given rise to a particular phylogenetic tree. Again, there is no room to inquire of the details of specific events, so we must find a way to think about the collection of all processes that have determined the shape of the tree. As in other historical sciences (such as archaeology and cosmology), we begin with a simple model that includes a rudimentary representation of the processes that create history, and we assess how well the model fits our picture of history. If the model fits poorly, then it is probably wrong and should be modified or replaced.

The general approach of comparing phylogeny estimates with the predictions of simple models of speciation and extinction was popularized by Raup et al. (1973) and Gould et al. (1977). These authors used models of random branching with damped equilibrium to generate phylogenetic trees and compared these with trees built from actual fossil data from a variety of taxa. Comparisons of diversity and clade shape revealed that many patterns in the fossil record were similar to those expected from random processes.

Recently, models have been developed that permit the study of speciation and extinction in the history of extant species (Hey 1992; Nee et al. 1994a,b). This approach offers two clear advantages over the use of fossils, at least in terms of data collection. First, the sampling of extant species can be completed to a high degree, without the presence of arbitrary gaps such as those caused by the vagaries of fossilization. This does not mean that sampling is either easy or thorough, and our knowledge of the diversity of extant taxa of large organisms is much better than for small organisms. Second, extant organisms have genotypic characters and a wider array of phenotypic characters than fossils, permitting better assessments of phylogenetic trees. However, data on extant taxa come from essentially a single point in time, and they offer no information akin to the procession of changes and the record of appearance and disappearance of forms that can be seen with a good fossil record.

We present a review of available methods for extant taxon studies, together with an evaluation of their utility. Two major points emerge from this analysis:

1. Phylogenetic tree models, and appropriate data, can reveal historical processes of speciation *and* extinction, even though the data come only from extant species.
2. The statistical power of this approach is limited, so that insights on extinction can be expected only when extinction has played a large role in the history of a group of extant species or when the species group is large.

We also apply several phylogenetic models to 11 data sets taken from the primary literature. These analyses conform, in general, to points 1 and 2. Most taxa reveal little evidence of extinction, with some informative exceptions.

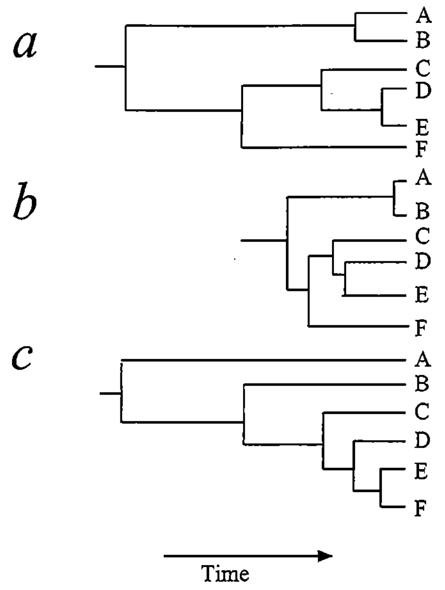

Figure 5.1. Hypothetical phylogenetic trees for groups of six species. Each species is indicated by an uppercase letter. Each tree consists of branches representing the persistence of species through time, and nodes (junctions of branches) that represent speciation, when one species splits into two. Three trees are shown. Both *a* and *b* have the same topology, but *b* has shorter distances between nodes. The topology of *c* is different, but here there are similar distances between nodes.

The Models

Phylogenies, hereafter often called trees, can be described as having two general attributes: length and shape (figure 5.1). The majority of this chapter focuses on questions about tree length, specifically within the context of rooted bifurcating trees, such as those in figure 5.1. These trees are called rooted because time has a direction such that one of the nodes is older than all of the others. Unrooted trees are not considered because they are not explicitly historical (i.e., time lacks direction). This review is also not concerned with multifurcating trees (i.e., more than two descendant species from a single node), primarily for simplicity and because the models that have been developed described bifurcating histories.

Almost without exception, the models of speciation and extinction that have been used to generate phylogeny predictions are Markov chains. By way of example, consider that at some point in time there is some number of species, N, and consider a sequence of short time intervals, within each of which there is some chance that speciation or extinction may occur. After

some random number of time intervals, t, a speciation event may occur or an extinction may occur. If speciation has occurred, then the number of species becomes $N + 1$; if extinction, then the number of species is, of course, $N - 1$. We may specify that this model has the Markov property, which means that the value of t does not depend in any way on the time since the previous speciation or extinction event. Models of this type are also usually homogeneous, meaning that the values of the extinction and speciation parameters are the same for each species and constant over time.

Discrete Markov chains have been widely used for the study of species turnover and taxon sizes (Raup et al. 1973; Gould et al. 1977; Stanley et al. 1981; Gilinsky and Good 1991). These models usually begin with a single species and then consider what can happen at successive discrete time intervals. In the very general model of Gilinsky and Good (1991, p. 150), at each time interval it is possible for each species to undergo speciation, or for a species to go extinct, or for nothing to happen. The important parameters under this model are the probability of speciation and the probability of extinction (the probability that neither occurs being 1 less the sum of these values). The principal shortcoming of these discrete models in a phylogenetic context is that time, per se, is not included. The random component of the model is the occurrence (or absence) of a speciation or extinction event within an interval. If the probabilities of speciation and extinction are low, then more intervals will pass without a change in N, on average. Thus, the number of intervals can be considered an indication of the passing of time, but the meaning of an interval remains unclear. Discrete Markov chains have a more appropriate application, for instance, in the study of the population dynamics of organisms that have discrete generations.

Continuous Markov chains, also widely used, differ from discrete models in that the time between events is an explicit random variable. Under a discrete model, the central question is whether or not a speciation or extinction event has occurred within an interval; under a continuous model, the central question is how much time has passed before the next speciation or extinction event occurs. In the parlance of continuous-time, homogenous Markov chains, speciation occurs at a specified rate (usually the Greek letter λ is used) and extinction occurs at a specified rate (usually μ). A discrete model with very low values for the speciation and extinction probabilities will generate results that are effectively indistinguishable from a continuous model. Continuous models are often more analytically tractable, relying on differential rather than difference equations. They are also usually easier to simulate, especially under low speciation and extinction rates, because they require fewer random numbers. Only continuous Markov chain models have been used to address questions on the structure of trees for extant species.

Tree Length

Perhaps the simplest model is one in which there is no extinction, but only speciation. This model, called a pure birth process or a Yule process, was first described in a phylogenetic context by Yule (1924) and was further developed for statistical tests on phylogenies by Hey (1992). Consider a single species that undergoes speciation at rate λ. Then the probability that the time until speciation takes place, time t, is given by

$$P(t) = \lambda e^{-\lambda t}. \qquad [1]$$

This is an exponential distribution, and it can be shown that only an exponential probability distribution fits a continuous homogeneous process under the Markov property. The expected value of t is $1/\lambda$, meaning that the average time until speciation is the inverse of the speciation rate. After speciation, there are two species. Because we are assuming that the speciation rate per species is constant, the time until the next speciation event is also an exponential distribution:

$$P(t_2) = 2\lambda e^{-2\lambda t_2}. \qquad [2]$$

On average, the time until the next speciation event is $1/(2\lambda)$, exactly half of the average time for the case of one species. At a time when there are N species, the time until the next speciation event is

$$P(t_N) = N\lambda e^{-N\lambda t_N}. \qquad [3]$$

In general, we see that the waiting time until the next event is less if there are more species that can undergo speciation.

Figure 5.2 provides an example of a phylogenetic tree under this model.In this example, the times between speciation events have been set to the expected values. This example reveals a phenomenon common to Markov chain phylogenetic models, as well as to phylogenetic trees estimated from actual data: the time between the deepest nodes of the tree (those farthest in the past) is greater than the times between the recent nodes of the tree.

A more complex model includes an extinction rate, μ. If there are N species, the probability that the time until either speciation or extinction occurs is given by

$$P(t_N) = N(\lambda + \mu)e^{-N(\lambda+\mu)t_N}. \qquad [4]$$

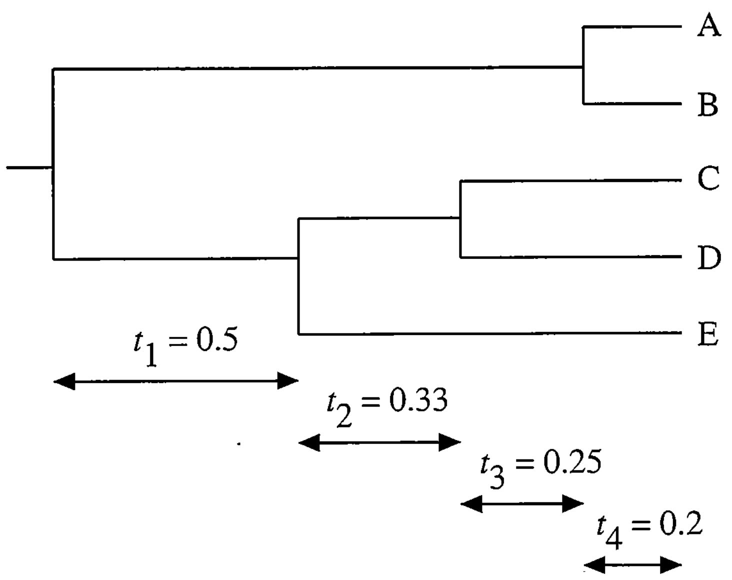

Figure 5.2. A hypothetical phylogenetic tree under a model of no extinction. The values of t represent the times between successive nodes and have been set to the expected values for a speciation rate of 1.0 per species. In general, there are $I + 1$ species persisting between node I and node $I + 1$, so the expected time between node I and node $I + 1$ is $1/(I + 1)$ (see text for further explanation).

The probability that the event is speciation is $\lambda/(\mu + \lambda)$ and the probability that it is extinction is one minus the probability of speciation, that is $\mu/(\mu + \lambda)$. This model, also called a pure birth and death process, has been employed in a phylogenetic context (Raup 1985; Nee et al. 1994b). Unlike the pure birth process, in which the number of species is equal to the number of events plus one, the number of species after some time of a birth and death process may have a wide distribution, and it will include the case of all species having gone extinct.

When extinction is included, an important distinction emerges between the true history of a set of species and the history that is perceived with a phylogenetic tree. A bifurcating phylogenetic tree of N species will contain exactly $N - 1$ nodes, each representing a case of speciation. However, if the true history includes speciation *and* extinction, then some of the speciation events will not be represented in the phylogenetic tree. Only those speciation events for which both daughter species are ancestors of the extant species at the tips of the tree will be represented by the phylogenetic tree (figure 5.3) (Hey 1992; Nee et al. 1992). In a recent series of papers, Nee and colleagues (1992, 1994a,b) have described the relationship between the true history of

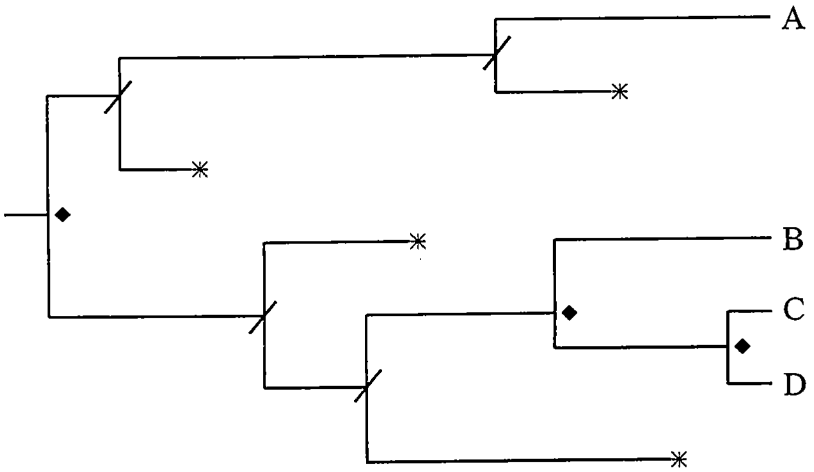

Figure 5.3. A hypothetical history of speciation and extinction. Four species (A, B, C, and D) are extant. Branches indicate all species that existed for some period of time. Only the four labeled branches (A through D) extend to the far right. Cases of extinction are indicated by a branch terminating in an asterisk (*). *Black diamonds* indicate those three nodes for which both descendant species were ancestors of extant species. These are the only speciation events that show up in a phylogenetic tree. Other speciation events are indicated with a *slash.*

speciation and extinction (which includes branches that end in extinction and do not reach the present) and the history perceived with a phylogeny estimated from extant species. If one considers the number of species present at times extending into the past, then this value will usually be lower (it cannot be greater) for the perceived history than for the actual history, because of extinction. Furthermore, this discrepancy is expected to be small for recent times and greater for times further in the past.

A special case of the birth and death process that is of special interest is when the speciation rate equals the extinction rate ($\lambda = \mu$). Like the pure birth model, this model is attractive for having just a single parameter. Also, most groups of organisms that have existed on earth have gone extinct (Raup 1994). This means that most groups of organisms have experienced identical levels of speciation and extinction (although the process may not have been homogeneous; see, e.g., Pearson 1992). A model in which speciation and extinction occur with equal rates is one tool to address the question of whether extant groups of organisms have similar processes of diversification and turnover as did extinct groups.

Yet another Markov chain model that has been employed within a phylogenetic context is one in which speciation and extinction always occur together (Hey 1992). Consider a group of N species, and suppose that whenever one undergoes speciation, another species becomes extinct, so that the number of species remains constant. This model was developed by Moran (1958) and is often used in a population genetic context for the case when population size is constant. It turns out that the waiting time between nodes in the perceived phylogeny of a group of N species under this model has a form that is very similar to those expected of a pure birth process. The probability distribution of the time between node I and node $I + 1$ is given by

$$P(t_i) = NB\gamma_i e^{-NB\gamma_i t_i}, \qquad [5]$$

where B is the speciation/extinction rate per species and γ_i is equal to $I(I - 1)/[N(N - 1)]$ (Kingman 1982; Tavare 1984; Hey 1992).

Hey (1992) developed the Yule model and the Moran model in parallel so that both could be compared with phylogenetic trees estimated from molecular data. In that paper, the Yule model was called model G, for growth, and the Moran model was called model C for constant. These various models can be arrayed as a sort of continuum with respect to the amount of extinction they include. Model G has no extinction; the general birth and death model has arbitrary rates for each process, but the case of equal rates is of special interest; and model C has speciation and extinction at equal rates, but these two processes always occur together.

To better compare the pure birth and death model with equal rates to models G and C, a computer simulation was developed. The simulation proceeded exactly as the birth and death process was described earlier. Furthermore, the number of species was recorded at the time of each event, as was the pattern of ancestor–descendant relationships that spanned events. Figure 5.4 shows the average time between successive nodes of phylogenetic trees under all three models. All three models predict increasingly greater time intervals further in the past, as one approaches the base of a phylogenetic tree.

Tree Shape

Two trees may have the same number of species, and identical time intervals between nodes, yet have very different topologies. In general, the time intervals between nodes can vary independently from the actual branching pattern. All of the models that have been described so far employ the assumption that the species that undergo speciation and extinction at any point in

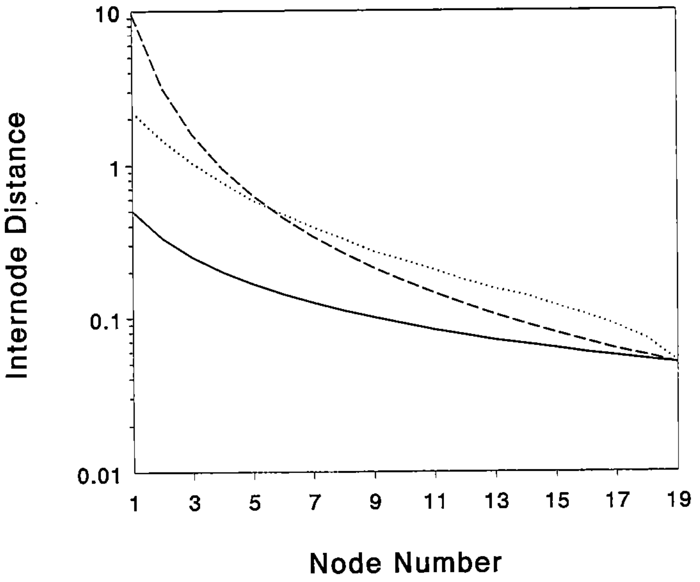

Figure 5.4. The expected distances between successive nodes for the phylogenetic tree of a monophyletic group of 20 species. Node 19 is the most recent; node 1 is the oldest. The speciation rate per species was set to 1.0 for models G (*solid line*) and C (*broken line*). The expected distance between node I and $I + 1$ for model G is equal to $1/(I + 1)$ and equal to $1/[I(I + 1)/19]$ for model C. For the model of equal but independent birth and death [$\lambda = \mu$ (*dotted line*)], both rates were set to 0.5 for each species (so that the time for node 19 would be equal to that for models G and C) and the internode distances were found via simulation. The simulations began with a single species and then proceeded as described in the text. Under this model, the number of species fluctuates randomly. To ensure that the results were not biased by recording the first trajectory that reached 20 species, the following protocol was followed. Each run included a large number of events (up to 500, although often all species became extinct prior to this time). If during a simulation, the number of species reached 20, then the simulation was used; if not, the simulation was discarded. During each run, all time points when the number of species was exactly 20 were recorded. After the conclusion of the run, one of these points was randomly selected and the history (which had been saved) was determined backwards from that point. The values indicated by the *dotted line* are the average internode distances found from 4000 successful runs.

time are randomly selected from those that exist at that point in time. This means that at a time when one of N species is undergoing speciation, there are N differently shaped histories that can result from that speciation event. If extinction is a separate possibility, then there are an additional distinct N number of trees as well. In general, the number of possible branching patterns of a phylogenetic tree is very large (Felsenstein 1978), and the value is even greater if nodes are ordered in time (Hey 1991), as they are in the models discussed here. For N equal to 5, there are 105 distinct rooted trees and 180 distinct rooted trees with ordered nodes. For N equal to 10, there are 3.45×10^7 distinct rooted trees and 2.57×10^9 distinct rooted trees with ordered nodes.

What kinds of tree shapes are predicted by the Markov chain models? To answer this question, it is first necessary to define *tree shape.* In the trivial case where tree shape means a topology with ordered nodes, then all possible outcomes are equally likely under the Markov chain models. However, tree shape may be defined in such a way that some shapes are more likely than others. The most common meaning of shape is balance or symmetry. Trees that are more balanced have roughly equal numbers of descendants extending from each node (e.g., compare figure 5.1a and figure 5.1c). Balance is most easily assessed for a single node, in which case the numbers of species descendant from the left and right sides of the node (L and R, respectively) are expected to be consistent with a random branching model (Slowinski and Guyer 1989; Sanderson and Donoghue 1994). However, for a node with M descendant species, such that $L + R = M$, all possible values of L ($L = 1$ to $M - 1$) are equally likely (Harding 1971; Slowinski 1990). This means that the models do not predict balance in the common vernacular of "symmetry," any more than they predict a lack of balance. It is still possible to ask whether the difference between L and R is bigger than expected by chance, although significant results are not possible without large values of N (Slowinski and Guyer 1989; Sanderson and Donoghue 1994). It is also possible to develop metrics for overall tree balance or symmetry, and several measures have been described (see Kirkpatrick and Slatkin 1993).

The specific assumption of the Markov chain models that is examined by tests of balance is the idea that all species are equally likely to undergo speciation or extinction. If there exist nodes such that the taxa on one side have different rates of speciation and/or extinction from those species on the other side, then the tree is expected to lack balance. A related but distinct method focuses specifically on whether the species that descend from each node are more or less likely than other species to be connected to the next node (Hey 1992). In effect, this test examines whether new species are more or less likely to undergo speciation than old species.

The remainder of this chapter is limited to tree length issues. The various Markov chain models generate predictions about lengths, yet they differ in the degree to which extinction is included and thus in the predicted pattern of variation. None of the tree shape analyses that have been developed are capable of distinguishing among the different Markov chain models.

The Data

In practice, fitting the Markov chain models to data on extant groups imposes significant practical constraints that limit the appropriateness of many data sets. One of the most problematic discrepancies between the models and comparative data is that the models generate predictions about time, whereas the data come in units of character change or divergence. Thus, to apply the models it must be roughly true that the amount of divergence that has occurred between species be proportional to the amount of time that has passed since their most recent common ancestral species. It is this requirement, that divergence be clocklike, that has the effect of limiting comparisons between models and data to those studies using molecular genetic data. For closely related species, the divergence of DNA is often roughly clocklike. Furthermore, the mutational process for DNA is fairly simplistic, relative to that for other traits, such that adjustments can be made to account for multiple mutations that can occur under high levels of divergence (see, e.g., Nei 1987). If divergence is clocklike, then the data are also very suitable for tests of tree shape. Any of a variety of methods can be used to generate accurate estimates of tree shape under these circumstances (Rohlf and Wooten 1988; Rohlf et al. 1990).

A second constraint arises from the fact that all of the models generate predictions about monophyletic groups: i.e., groups for which all of the extant descendants of the most ancestral species (represented by the basal node) are included. It is necessary, then, that data sets used for testing models also completely represent monophyletic groups. Model C, in which speciation and extinction are coupled, is a partial exception. It turns out that equation 5 applies for a random sample of species from a monophyletic group. At any rate, the requirement of monophyly presents three difficulties. First, monophyly is a statement about phylogenetic history, so that the monophyletic status of a group of species selected for phylogenetic study is usually not well known prior to the collection of data. Second, the requirement that all species of a monophyletic group be included has the effect of ruling out groups of organisms for which thorough global diversity surveys have not been made. Last, many phylogenetic studies that employ molecular genetic methods, even those focused on small taxonomic groups, have not included all of the known species for the taxon under investigation.

Testing Models

In practice, the conditions of monophyly and clocklike evolution are likely to be approximated only for relatively recent taxa containing small numbers of species. Typically, the monophyletic status of small groups is more likely to be known, because they are more often studied as part of a larger group. Also, recently diverged species are more likely to accumulate mutations at similar rates, so that the assumption of clocklike divergence is more likely to fit for recent groups than for older species groups.

It is also important to emphasize that the models that have been described do not incorporate the uncertainty that comes with actual data sets. In particular, phylogenetic tree estimates typically have considerable uncertainty associated with branch lengths and with topology. Similarly, the models assume monophyly, but in practice this is difficult to know with certainty. In short, the models ignore sources of variation that are inherent to phylogenetic estimates.

A thorough literature search has returned 11 data sets for which DNA divergence data have been estimated from all of the species of a taxon. Six of these data sets were included in Hey (1992), with the remainder having appeared or been found since then. Three types of DNA studies have been included: studies on restriction fragment length polymorphism (RFLP), DNA-DNA hybridization studies, and DNA sequencing studies.

Table 5.1 presents the results of several analyses on these data sets. For each data set, models G and C were applied, and for each of these models, the likelihood ratio statistic, Λ, and the probability of the likelihood ratio statistic, p, are presented. Also shown are the estimates of the speciation rate for model G, $\hat{a}$ (this notation is in keeping with Hey 1992, although this parameter is a speciation rate similar to λ), the speciation/extinction parameter for model C, $\hat{b}$, and the 95% confidence intervals of these estimates. The model of Nee et al. (1994b) was also applied to these data. Expression 21 of Nee et al. (1994b) was maximized, under the data, to generate estimates of λ and μ, and these are also given in table 5.1.

An error was found in Hey (1992). In that paper, it was reported that the most recent node should not be included in the analyses for models G and C. In fact, this node is suitable, and the arguments against using it are a fine example of the Bus Stop fallacy (Feller 1971, p.12). Correcting this error causes some small changes in Table 5.1 relative to the analyses on the same data sets reported in Hey (1992).

In Hey (1992), the principal finding from an analysis similar to that presented in Table 5.1 was that model G (no speciation) fit the data better than model C. This pattern persists with the addition of more species groups; 10 of the 11 data sets generate a lower likelihood ratio statistic under model G than under model C (and concomitantly, a higher value of p). A striking parallel to this pattern is that in all ten of these cases, the estimate of μ, under the Nee et al. model, is zero and the estimate of λ is very similar to $\hat{a}$. It

TABLE 5.1 Fitting models to data

			Model G				Model C				Independent Birth and Death[g]	
Data set	Data type[a]	N[b]	$\hat{a}$[c]	95% CI limit[d]	Λ[e]	p[f]	b[c]	95% CI limit	Λ	p	λ	μ
Subgroup *melanogaster*	DNA-DNA	8	0.59	0.31–1.4	3.25	0.91	0.13	0.07–0.33	7.04	0.54	0.63	0.0
Tribe Hadenoecini	DNA-DNA	9	0.26	0.15–0.63	4.94	0.75	0.04	0.02–0.10	12.6	0.15	0.26	0.0
Family Gruidae	DNA-DNA	14	1.52	0.95–2.8	13.5	0.49	0.23	0.14–0.43	17.6	0.22	1.79	0.0
Genus *Equus*	RFLP	6	0.29	0.14–0.85	4.76	0.41	0.07	0.03–0.21	9.59	0.09	0.25	0.0
Sect. *Peripetasma*	RFLP	9	2.27	1.27–5.4	5.4	0.71	0.42	0.23–0.96	11.2	0.21	2.38	0.0
Subgroup *nasuta*	RFLP	10	0.33	0.19–0.73	5.08	0.83	0.07	0.04–0.16	7.82	0.58	0.39	0.0
Complex Narbonensis	RFLP	6	0.37	0.18–1.17	9.75	0.08	0.10	0.05–0.30	13.4	0.02*	0.38	0.0
Genus *Krigia*	RFLP	7	2.0	1.0–5.3	2.36	0.85	0.66	0.34–1.8	2.55	0.83	2.3	0.0
Genus *Graptemys*	DNAseq	13	0.87	0.53–1.66	13.0	0.43	0.13	0.08–0.25	18.6	0.13	0.97	0.0
Genus *Microseris*	RFLP	8	4.14	2.23–10.3	13.8	0.06	1.91	1.0–4.8	7.55	0.37	13.9	11.9
Genus *Calycadenia*	DNAseq	12	0.55	0.33–1.1	6.4	0.86	0.08	0.05–0.16	14.0	0.29	0.61	0.0

[a] DNA-DNA, DNA hybridization; RFLP (restriction fragment length polymorphism), restriction enzyme assays of variation; DNAseq, DNA sequencing studies.

[b] N, number of species in the data set. The number of internode distances used for models G and C is equal to $N - 1$.

[c] Values of $\hat{a}$ and $\hat{b}$ were calculated using expressions 3 and 10 of Hey (1992), respectively, with the exception that all $N - 1$ internode distances were used.

[d] The 95 percent confidence interval (CI) limits of $\hat{a}$ were calculated as follows: Using the observed value of $\hat{a}$, 5000 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 percent and 2.5 percent positions in the ranking, respectively. The same procedure was done for the 95 percent CI limits of $\hat{b}$.

[e] Λ is the likelihood ratio statistic, generated using expression (13) of Hey (1992), again with the exception that all $N - 1$ internodes were used.

[f] p is the probability of getting a likelihood ratio statistic equal to or greater than that observed under the model; it was generated using the same simulations generated for the CI intervals.

[g] λ and μ are the birth and death rate parameter estimates under the independent birth and death model and were calculated using expression 21 of Nee et al. (1994). The internode distances required for the calculations of all rate parameters were generated by applying the UPGMA protocol in the SAHN program of NTSYS (Rohlf 1985) to the distance data reported in the references.

[h] Subgroup *melanogaster* (Caccone et al. 1988): 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.*

Complex Narbonensis of genus *Vicia* (Raina and Ogihara 1994): chloroplast sequence diversity based on restriction site analysis on a six-member group of legumes.

Genus *Krigia* (Kim et al. 1992): chloroplast sequence divergence based on restriction site analysis of seven species of dwarf dandelions.

Genus *Graptemys* (Lamb et al. 1994): combined cytochrome *b* and control region sequences from the mitochondria of 13 members of this genus of turtles.

Genus *Microseris* (Asteraceae) (Wallace and Jansen 1990): chloroplast sequence divergence based on restriction site analysis. The full data set included all 16 species of the genus, but several species exhibited no net interspecific variation. These species were collapsed on one another, with the effect that recent internode distances of length 0 were ignored.

Genus *Calycadenia* (Compositae) (Baldwin 1993): DNA sequences of nuclear ribosomal gene spacers. The full data set included all 11 species, but some exhibited no net divergence. These species were collapsed on one another, with the effect that recent internode distances of length 0 were ignored.

seems that the predominant history among small monophyletic groups of extant species includes speciation but not extinction. The one exception to this pattern, the data for *Microseris*, is better fit by model C than model G, and it generates high and similar estimates of λ and μ.

However, before accepting the conclusion of zero or low extinction, it is important to inquire of the sensitivity of these tests to the presence of extinction. There are three distinct causes for the concern that extinction has not been well measured: insensitive tests, sampling bias, and estimator bias. These topics are discussed in turn.

Insensitive tests: It has been shown that a simple model of history without extinction is expected to have increasingly greater node intervals as one looks deeper in a phylogenetic tree. This pattern is also expected of histories with extinction, only more so. The difficulty is that a history of equal and independent processes of speciation *and* extinction does not have a pattern of node intervals that differs sharply from the case of no extinction. In figure 5.4, the shapes of the curves under model G and the independent birth and death model are very similar. Figure 5.4 suggests that a history of extinction and speciation might generate phylogenetic trees similar to those with only speciation, albeit with lower apparent rates of speciation.

Sample bias: All of the Markov chain models that have been described are intended to approximate the history of extant species groups, and these groups necessarily have experienced more cases of speciation than extinction—a group of *N* species has had exactly *N* more cases of speciation than extinction. However, it is possible that the underlying speciation and extinction rates are no different for extant groups than for extinct groups (which have equal amounts of both). Thus, when we study a group at a time prior to the extinction of all species within the group, we have selected a history that may underrepresent any underlying propensity to extinction. This condition is realized in the method of Nee et al. (1994b), which requires that the estimate of the speciation rate be greater than that for the extinction rate ($\lambda > \mu$).

Estimator bias: It is possible that there is an actual statistical bias in the estimators. In other words, the expected value of a parameter estimate may differ from the true value of the parameter, even when samples are not biased and when evolution has occurred as assumed and the model is completely correct. The estimates $\hat{a}$ and $\hat{b}$ in table 5.1 have a slight bias of this type (Hey 1992).

To examine the difficulty of distinguishing extinction, we have employed a simulation in which birth and death rates are equal. In some respects, this is the most attractive null model for tests of the structure of phylogenetic trees: the model has only a single parameter, and it enjoys the justification that birth and death rates have been equal for most groups of organism (those that have gone extinct). We used the model of Nee et al. (1994b) to inquire of the bias in estimates of the parameters λ and μ. Figure 5.5 shows the average estimates of λ and μ for a wide range of differently sized monophyletic groups. The simulations were run with $\lambda = \mu = 0.5$, with 1000 replicates per sample size. It is clear that both parameters are highly biased, even for species groups as large as 40. It is expected that most of this bias is the result of sampling bias, and is not estimator bias. Although λ and μ are estimated by maximum likelihood, a method that often generates biased estimates, the similar values found for λ and $\hat{a}$ in table 5.1 argue against a high estimator bias for λ. It is known that the estimator bias for $\hat{a}$ is slight (Hey 1992).

We next asked about the statistical power of these methods. Statistical power is the probability of making the correct decision to reject the null model when the null model is false and the alternative model is true. To assess power, it is also necessary to specify an alternative model. Because our concern is that these methods may have little power to detect extinction, we

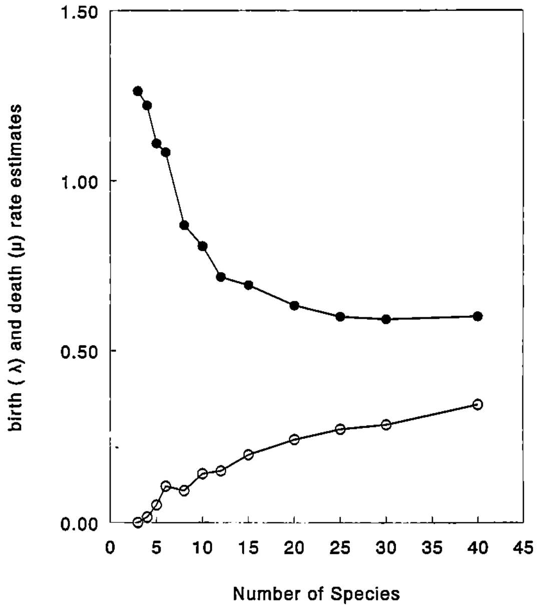

Figure 5.5. Assessing the bias of estimates of speciation and extinction. Average values, from 4000 simulations, are shown for λ (*filled circles*) and μ (*open circles*), estimated by the method of Nee et al. 1994b. For both λ and μ, the true values were set to 0.5. Simulations were conducted under the model of independent speciation and extinction as described in the text and in the figure 5.4 legend.

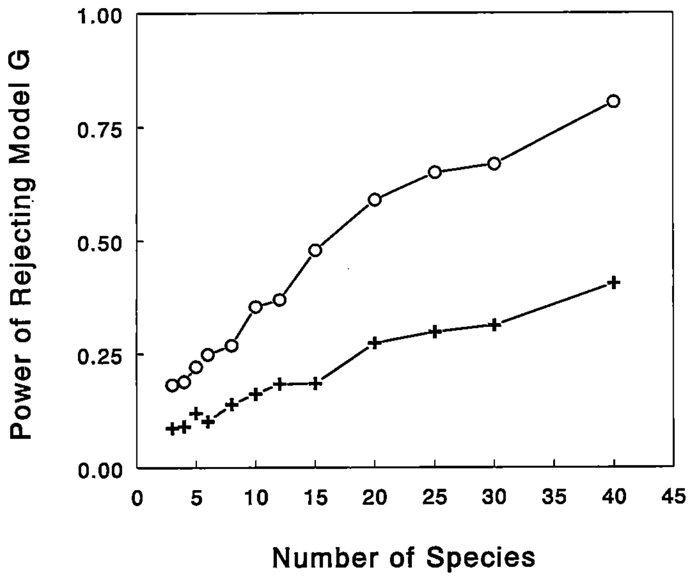

Figure 5.6. Statistical power of tests of model G. Data sets were simulated under model C (*open circles*) and the model of equal but independent birth and death rates [$\lambda = \mu$ (*pluses*)]. For each simulated data set, model G was applied and the likelihood ratio test (Hey 1992) used to assess whether model G could be rejected. Each point represents the proportion of 1000 independent simulated data sets that were rejected, under model G, at the 95 percent level.

will use the growth model (no extinction) as the null hypothesis, and models that include extinction as alternative hypotheses. In other words, we simulated phylogenetic trees under a model that includes extinction (i.e., the alternative hypothesis is true) and then fit the growth model to the simulated data and asked if the growth model was rejected. Two alternative hypotheses were considered: the independent birth and death model with $\lambda = \mu$, and the constant model with complete linkage of speciation and extinction. The results over a range of sample sizes are given in figure 5.6.

As expected, power increases with sample size. However, power is quite low under both alternative models for sample sizes in the range of those reported in table 5. 1. Perhaps of greatest concern is that the model of independent but equal speciation and extinction processes has very low power across the entire range of sample sizes considered. In short, if speciation and extinction have been going on independently under a homogeneous Markov process, then most of the phylogenies are going to be found consistent with a model that lacks extinction. The phylogenies will certainly be deeper with the extinction than without, but the spacing of the nodes will not usually depart much from the pattern expected without extinction. If this is the case, then the estimated speciation rate, assuming no extinction, will be lower than the true rate.

Although most species groups in table 5.1 exhibit patterns that are apparently well fit by model G, consistent with a lack of extinction, there are some noteworthy departures. One taxon [genus *Microseris* (family Asteraceae, tribe Lactuceae)] is fit poorly by model G and better by model C, and it has high and similar estimates for λ and μ. The estimate of λ (13.9) is over 3 times that for $\hat{a}$ (4.14). The poor fit of model G, and the discrepancy between the parameters $\hat{a}$ and λ, suggest a history with frequent speciation and nearly as much extinction. The final phylogenetic tree suggests a much lower rate of speciation when fit to model G, but the fit is poor. The UPGMA tree for this species group is shown in figure 5.7*a*, and it can be seen that the intervals between the deepest nodes are much greater than those for the shallowest nodes. A second taxon that does not fit the overall pattern is the Narbonensis complex in the legume genus *Vicia.* Both model G and C fit poorly, and model C is rejected (table 5.1). Figure 5.7b shows a UPGMA tree for this group, in which it can be seen that there is a pattern of short internodes near the base of the tree and near the tips of the tree, with the longest internodes in the middle. This type of pattern is not predicted by any of the models presented here.

Conclusions

Extinction and speciation are the yin and yang of life's diversification. The current state of the web of life that occurs on our planet can be viewed as a function of those organisms that have persisted or in terms of those organisms that have perished. However, biologists and paleontologists tend to overlook extinction. Despite some initial emphasis by Darwin, "extinction largely dropped out of the consciousness of evolutionary biologists and paleobiologists" (Raup 1994). Without dwelling on the psyche of scientists, some of this "extinction gap" can be explained by the differences between extinct and extant organisms. For biologists who study diversity at the present moment, or for paleobiologists who study diversity at time points in the past, information on those organisms that exist at a specific point in time is more accessible than for those that are absent. The sampling bias described in this chapter for studies on extant species is essentially a manifestation of the difference in accessibility between extinct and extant species. In fact, extinction is expected to leave a mark on the phylogenetic history of those organisms that have not gone extinct, and this chapter has reviewed several models that can be used to assess the role of extinction. Unfortunately, the

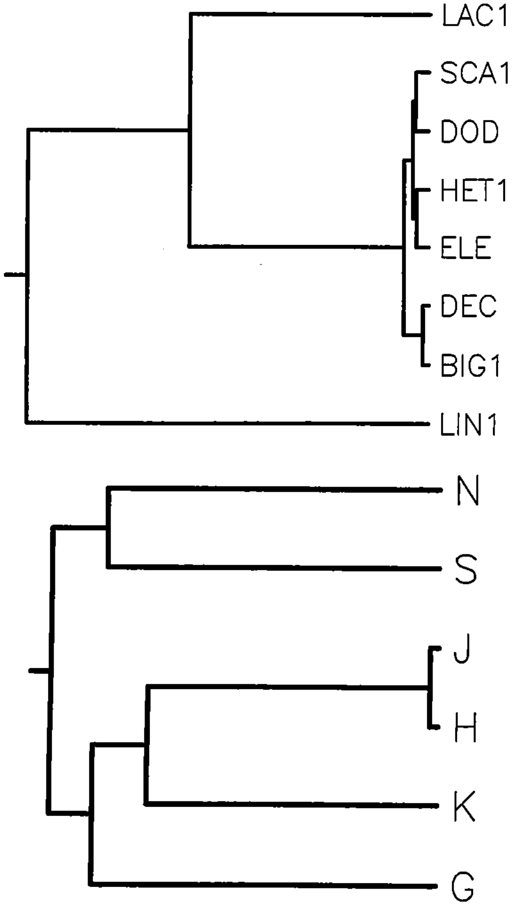

Figure 5.7. UPGMA trees (Sneath and Sokal 1973). **(A)** Genus *Microseris* (Asteraceae) (Wallace and Jensen 1990). Data presented in table 2 of Wallace and Jensen (1990) were converted to estimates of nucleotide substitutions and adjusted for multiple hits (Nei and Li 1979), and they were used to generate the distance matrix from which the UPGMA tree was made. The full data set included 16 species, but several recently diverged species exhibited no net interspecific variation (see table 5.1 legend). **(B)** The Narbonensis complex of genus *Vicia* (Raina and Ogihara 1994). The data presented in table 3 of Raina and Ogihara (1994) were first converted to estimates of the numbers of nucleotide substitutions and adjusted for multiple hits (Nei and Li 1979).

mark that extinction leaves is subtle, and usually there is little statistical power to distinguish between models that include extinction and those that do not. Extinction in a phylogenetic history causes longer internode distances deep in a tree, but it can be difficult to distinguish between a model with extinction, and a model without extinction at a lower rate of speciation.

Our analysis of 11 data sets has revealed some interesting patterns. First, 10 out of the 11 data sets fit the growth model best and return an estimated extinction rate of zero. Is this surprising? The species groups listed in table 5.1 share the traits of being small and young (typically less than 10 million years since the last common ancestor of the group). It is possible that extinction has not been common in these groups. This conclusion necessarily entails a corollary that extinction is more common among older groups of species. At present, we cannot say if the apparent absence of extinction is because of a true lack of extinction or whether it is caused by sampling bias, and the lack of resolution with small species groups, with respect to the tested models.

The data from *Microseris,* a genus of annual and perennial herbs (Asteraceae: Lactuceae) (Wallace and Jansen 1990) is consistent with a history that includes extinction. This exception to the general pattern suggests that we are able to detect extinction at least in some cases and that not all small extant groups are in a phase of growth without extinction. Another exception is the Narbonensis complex, a group of species that are thought to be the closest wild relatives of the broad bean (*Vicia faba*) (Raina and Oghihara 1994). These species have a history that did not closely match any of the models tested. In short, the longest internodes are not deep in the tree, but rather at an intermediate position, as if caused by some non-homogeneous process of speciation and/or extinction.

The ability to detect speciation and extinction with phylogenetic trees expands as the number of species within a monophyletic taxa increases. The current bloom in systematic research using DNA tools should increase the number of appropriate data sets and their size. Given some, albeit limited, resolution of speciation rates and extinction, we can look forward to larger data sets and the research questions that may become accessible.

Some concepts discussed elsewhere in this volume suggest a few a priori distinctions that could become interesting contrasts as more data become available:

1. Good versus poor dispersers. As suggested by Harrison (chapter 2), if taxa exist as metapopulations, then well-dispersing taxa should persist longer than taxa that disperse poorly.

2. Abundant versus rare species. Abundant species with large ranges should persist longer and have more chances to speciate (Maurer and Nott, chapter 3).

One question of special interest from our findings would be a contrast between older versus younger clades (focusing on the time in the past of the oldest node). As suggested by our analyses, very recent groups may be in growth-only mode, while older species may be holdouts in shrinking groups.

6

Dynamics of Diversification in State Space

Daniel W. McShea

America spread west in the mid to late nineteenth century. In our own century, the story of the mass migration has been told and retold in countless ballads and novels. And in now-classic movie westerns, it has been epitomized in scenes of long wagon trains winding across the prairie toward the setting sun.

Census data from that period confirm the demographic aspect of the story: the mean location of Americans—the country's center of gravity, so to speak—shifted westward. However, the data also show that from two broad belts occupying the middle third of the country, the west-central states and the mountain states, about the same number of people moved *east* as moved west, at least during the first part of the migration from 1850 to 1880 (data from Mitchell 1983). Actually, slightly *more* moved east during this period.

The pattern is puzzling at first. The country as a whole spread west, while throughout the West, slightly more people migrated east. But it is consistent with a simple explanation. The westward shift of the mean could be explained by short, local westward movements by people in the East; unlike other regions of the country, migration there is necessarily biased westward, because migration eastward is blocked by the Atlantic Ocean. The principle is illustrated abstractly in figure 6.1A. The horizontal axis is location in space, reduced to a single dimension, east and west, and the vertical axis is time. Suppose for simplicity that we are dealing with an asexually reproducing spe-