Paleobiology, 30(2), 2004, pp. 172–178

MATTERS OF THE RECORD

Extinct meets extant: simple models in paleontology and molecular phylogenetics

Sean Nee

Ashworth Laboratories, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom. E-mail: sean.nee@ed.ac.uk

Accepted: 9 November 2003

Introduction

Paleontologists have a long tradition of the use of mathematical models to assist in describing and understanding patterns of diversification through time (e.g., Raup et al. 1973; Stanley 1975; Sepkoski 1978; Raup 1985; Foote 1988; Gilinsky and Good 1989). This is natural, as the information, phylogenetic and otherwise, that paleontologists work with comes equipped with a temporal dimension, albeit approximate, which endows these phylogenies with information about the tempo of evolution as well as the genealogical relationships among the lineages. Mathematical and statistical modeling are the tools for unlocking the quantitative information in the phylogenies.

Recently, molecular phylogenetics (e.g., Hillis et al. 1996) has created a new source of phylogenies with a temporal dimension, now frequently provided by molecular clocks. Many people have applied mathematical models to these phylogenies as well. Here, I highlight the areas of overlap as well as differences in what the simple models in the two fields have to tell us. To save ink, I will hereafter refer to paleontology as P and molecular phylogenetics as MP.

First I review the use of simple mathematical models to extract information about the tempo of evolution from phylogenies in the fields of P and MP. The same models, or variants thereof—these being the birth, birth-death, and Moran models—are used in the two areas, but there are differences in what they tell us, arising from differences in the nature of the phylogenies themselves. Finally, I address a high-profile assault on this common framework of understanding that has recently been launched.

The Pure Birth Process

This is one of the two simplest mathematical models used in P and MP (the other is the Moran process—discussed below) and, in its stochastic form, was one of the first stochastic processes to have been studied (Yule 1924; Kendall 1948, 1949; Feller 1957; Bailey 1964). It assumes that clades grow as follows. At each point in time, each lineage (or higher taxon) has the same probability, b, as every other to give birth to a new lineage, and extinction—or lineage death—does not occur. Under this model, average clade size, $N(t)$, grows exponentially: $N(t) = N(0)e^{bt}$. This deterministic result from the pure birth process has been used by Sepkoski (e.g., 1978) and Stanley (e.g., 1975) among others.

It is notable that the statistician who first studied this process, Yule (1924), was inspired to do so by exactly the same sorts of questions that motivate us. The Willis in his paper's title (''A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, FRS'') was an expert on angiosperms who had formed what were perceived to be anti-Darwinian views on the sizes and distributions of species and higher taxa. The paper contains a thorough treatment of the pure birth process and even has such contemporary features as the use of the birth process to estimate the rate of cladogenesis of the angiosperms. Happily, the birth process is often called the Yule process, anchoring its origins in our own areas of interest. For a time, a competing name was the Furry process, after the astronomer who stud-

 0094-8373/04/3002-0002/$1.00

ied it in the context of cosmic showers (e.g., Kendall 1949), but this failed to catch on.

Baldwin and Sanderson (1998) provide a contemporary example of such a study in MP; they find that the Hawaiian silversword alliance, a group of 28 plant species exhibiting enormous morphological variation had a remarkably high speciation rate (0.56 ± 0.17 spp./Myr), exceeding the average rates of continental radiations. (As it happens, the confidence intervals they reported are wider than they need be [Nee 2001].) Compare, for example, the balanoid barnacles, with a rate of 0.12 spp./Myr, which is itself higher than a mean rate of 0.07 for a variety of bivalve and gastropod molluscs—a P study (Stanley and Newman 1980); or a rate of 0.342 spp./Myr for the Old World monkeys, the most rapidly radiating primate clade—an MP study (Purvis et al. 1995).

The birth process assumes speciation is instantaneous. Nowadays, we see this as an abstraction, or an approximation, that makes for a tractable model. Interestingly, this was not Yule's viewpoint. He treated it as a biological postulate: Willis took a mutationist/saltationist/''hopeful monster'' view of speciation. Strangely enough, one recent extension of the birth process has been to relax the assumption of instantaneous speciation (Losos and Adler 1995). This has the unfortunate consequence that the model can no longer be studied analytically, relying entirely on simulation.

If clades grow in accord with this birth process model, then a semilogarithmic plot of the size of the clade over time should appear linear and the slope of the plot provides an estimate of the rate of cladogenesis. Sepkoski (1978: Fig. 3) provides what must be the best example from P. The increase in the numbers of metazoan orders over the Vendian and Lower Cambrian is almost *too* linear, with the linear regression $r^2 = 0.994$! Nee et al. (1995) provide an example from MP with somewhat more stochastic wobble—the growth of the *Drosophila melanogaster* subgroup (Fig. 1).

The slope of the semilog plot is a perfectly good estimator of the rate of growth of the clade even if just fit by eye. Of course, we could instead use maximum likelihood procedures to estimate the rate taking, for example, the time intervals between the appearances of new lineages as the raw data. However, it turns out that putting a confidence interval on the rate turns out to be remarkably unstraightforward. This was first noted in MP by Baldwin and Sanderson (1998) and received a thorough treatment by Nee (2001). It was surprising that such a simple probability model would not readily yield a confidence interval.

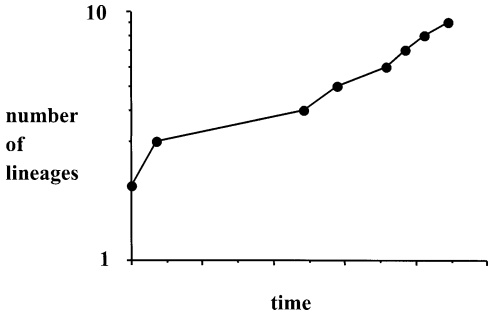

FIGURE 1. Semilogarithmic plot showing the increase over time in the number of lineages in the phylogeny of the *Drosophila melanogaster* subgroup constructed by Caccone et al. (1988). The timescale is not shown. In general, the timescale is in units of genetic distance or, if the molecular clock has been calibrated, actual time. The linearity of this plot shows that the growth of this clade accords with the pure birth process.

Before moving on to a natural extension of the pure birth model, I want to note the following. Although simple, I argued in an earlier paper (Nee 2001) that the birth process model is, in fact, one of the two most important models in this area. This is because a multitude of questions that may be of primary interest can be recast in the form ''is the pure birth model appropriate for my data?'' Examples of such questions are Has the rate of cladogenesis changed over time? Do the rates differ in these different clades? It so happens that the pure birth model has features that allow an arsenal of preexisting statistical tests to be brought to bear on such questions. The second most important model is the Moran model (discussed below).

The Birth-Death Process

A natural extension of the pure birth process is to include a constant probability of extinction, i.e., death, d, for each lineage (or higher taxon) at each point in time. For P, this introduces no radically new features. Clades

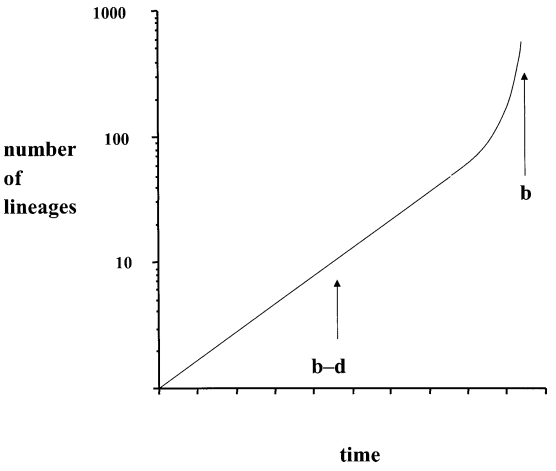

FIGURE 2. Theoretically expected increase in the number of lineages in a molecular phylogeny over time that has grown according to a birth-death process. Over much of the history, the slope of the semilogarithmic plot is expected to be the net rate of cladogenesis, $b - d$, and this slope is expected to asymptotically approach the speciation rate b as we approach the present. This is a visual illustration of the fact that it is, in principle, possible to estimate separately speciation and extinction rates from molecular phylogenies.

still grow exponentially, with average clade size increasing as follows: $N(t) = N(0)e^{(b-d)t}$. In fact, researchers such as Sepkoski (1978) and Stanley (1975) have always had this model in mind, referring to net rate of growth or cladogenesis, $b - d$, of clades. However, because only a single quantity, net growth rate, is involved I found it convenient to introduce these authors in the previous section.

For MP, however, the introduction of extinction makes an enormous difference. This is simply because the phylogenies that are produced by MP are based on extant species: hence, only those lineages that have at least one descendant at the present day are represented in the phylogeny. The effect of this is to cause an apparent increase in the rate of cladogenesis as we approach the present (Fig. 2) simply because species that have arisen recently have had less time to go extinct. This effect is larger the larger the ratio d/b. As discussed using the example of salamanders (Nee et al. 1995) one could misconstrue as rapidly improving health of the clade what is, in fact, considerable proneness to extinction.

Figure 2 provides a visual suggestion that, surprisingly, it may be possible to estimate speciation and extinction rates separately from molecular phylogenetic data, even though, perforce, they do not contain any explicit information about extinct lineages. This is, in fact, the case, and examples can be found in Nee et al. 1995 and Purvis et al. 1995. The ability to do this in MP is closely related to the fact that, in P, it is possible to estimate b and d separately from data on the survivorship of higher taxa, such as genera (Raup 1975, 1985; Foote 1988). What is the connection? In MP, a lineage that arose at some point in time, t, in the past will appear in the molecular phylogeny only if it has at least one descendant at the present day. In P, a genus that arose at some point in the past has a specified probability of surviving to a later time, t, i.e., having at least one descendant at the later time t. These probabilities, $\Pr(t)$, are the same:

$$\Pr(t) = \frac{b - d}{b - de^{-(b-d)t}}$$

So, the same probability expression appears in (a) the stochastic theory for the growth of molecular phylogenies and (b) the stochastic theory for the survivorship of higher taxa in the fossil record. This theory was used by Foote (1988) to estimate a remarkably high speciation rate of 0.4 spp./Myr for Cambrian trilobites. It should be noted, however, that in both P and MP the power to resolve birth and death rates separately is substantially lower than the power to estimate the composite net diversification rate, i.e., birth rate minus death rate. This is illustrated in MP by Nee et al. (1995).

Broadly speaking, there are two perspectives one can take on the phylogenetic data to be used for parameter estimation. We can look at the time intervals between successive nodes in the phylogeny, i.e., the rate at which new taxa appear, as our raw data. This is the approach taken in the previous section by authors examining semilogarithmic plots to estimate birth rates and can be used to estimate both origination and extinction rates in MP (e.g., Nee and May 1994). Or we can ''break'' the tree into its component branches and examine their lengths. This approach is familiar to most paleontologists as cohort survivorship analysis (e.g., Raup 1985; Foote 2001), which is used to estimate extinction rates, as well as

to explore any departures from null models such as the hypothesis of rate constancy. It can also be used to estimate origination rates by studying ''prenascence'' curves, i.e., the origination times of the members of a cohort existing at a particular time (Foote 2001); this is simply a time-reversed survivorship analysis. This approach has been suggested for MP by Paradis (1997), the members of a clade alive today defining a natural cohort, and this perspective on the tree was, in fact, also used by Nee and May (1994).

The Moran Process

The previous models are suitable for radiations that have not hit any limits to diversity. At the other extreme, we require a model for clades that have reached a plateau, as, for example, metazoan orders appear to have done in the Ordovician and Silurian (Sepkoski 1978). A useful model for this was first introduced in population genetics by Moran (1958): at each point in time, each lineage (or higher taxon) has a probability of going extinct, and when a lineage does go extinct it is replaced by the progeny of another lineage chosen at random. So in this model the clade is kept at a constant size deterministically. There are many analytical results available for this process in the population genetics literature, particularly that branch known as ''coalescence theory'' (e.g., Hudson 1990), and these tell us what a molecular phylogeny of a clade that has grown according to this process should look like.

This process does not differ substantially from the simulation algorithm studied by Raup et al. (1973), nor from the algorithm used by Sepkoski (1978) and Sepkoski and Kendrick (1993), once the plateau has been reached for some time. (In Sepkoski 1978 the clade grows logistically to the plateau.) In the simulations of Raup et al. (1973), they decided on a ceiling diversity and set $b = d$ at this ceiling. When the diversity dropped below the ceiling, they set $b > d$ to get it back there; similarly, they set $b < d$ when diversity rises above. There is a *big* difference between this model and a birth-death process in which $b = d$ and these rates are kept constant: in the latter model clade extinction is inevitable, for example. The latter model has been studied extensively in P (Foote 1988; Kitchell and MacLeod 1988; Uhen 1996) and not at all, to my knowledge, in MP. This is understandable: in P, which sees clades coming and going over time, the model is a natural one; in MP, we only see clades coming. The Moran model was introduced into MP by Hey (1992), although not identified as such.

Again, there is a big difference in the behavior of this model between P and MP, arising from the fact that MP does not ''see'' extinct lineages. As might be expected from the behavior of the birth-death process, it produces an apparently rapidly accelerating rate of cladogenesis as we approach the present. But an important difference between the Moran model and the birth-death process model is that the Moran, like the Yule model, readily opens up a preexisting arsenal of statistical tests that one might want to use, such as tests of the adequacy of the model given the data, parameter estimation, or comparisons of the parameters of different clades. The reason for this is as follows. Taking the time between the nodes in the phylogeny as our data, under both the Yule and the Moran processes there are simple transformations of the data that transform them into i.i.d. exponential variables (Nee 2001). This turns the temporal history of the clade into a Poisson process, and such processes have been studied for years by statisticians (Cox and Lewis 1966). For this reason, I have suggested (Nee 2001) that this model and the Yule process be the two tools of choice in the investigator's tool box for MP. They have, of course, been tools of choice in P for years.

Numbers of Subtaxa per Taxon

In a clade, how many families have one genus, how many have two genera, how many have three, etc.? It has been known since the work of Willis in the 1920s that the frequency distributions answering such questions (or the numbers of families in orders, etc.) are of the ''hollow curve'' variety, with a mode of monotypic taxa and a long tail (e.g., Sepkoski 1978: Fig. 11). Sepkoski (1978) was interested in such distributions in the context of the question of whether or not studies of diversity that

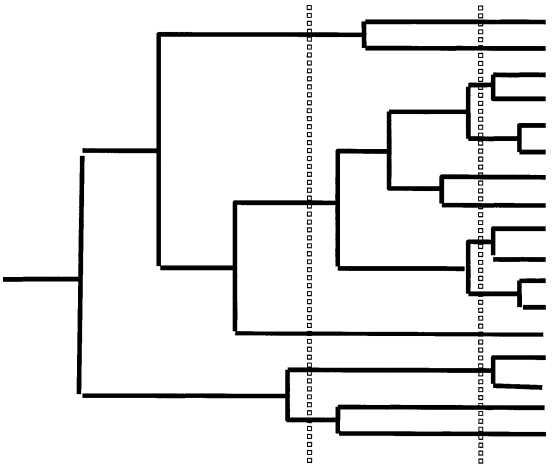

FIGURE 3. Illustration of the subtaxa per taxon analysis discussed in the text. For each lineage crossing the left vertical line, how many daughter lineages does it have by the time we reach the right vertical line? As this is a hypothetical molecular phylogeny, there are no extinct lineages.

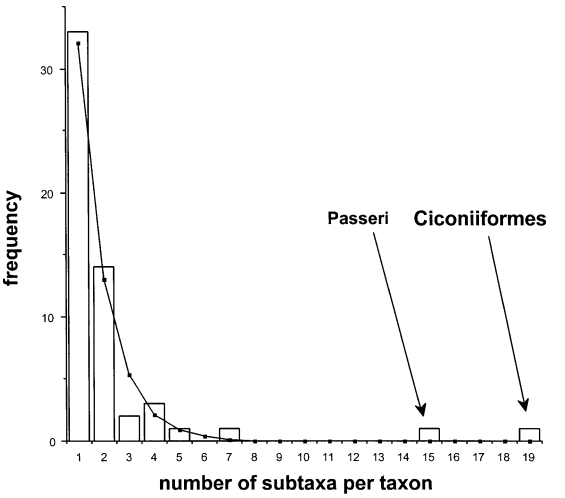

FIGURE 4. The results of applying the analysis illustrated in Figure 3 to Sibley and Ahlquist's molecular phylogeny of the birds. The line is a fitted geometric distribution excluding the two indicated statistical outliers.

are only resolved to, say, ordinal level are informative about diversification at the species level. He concluded that they are.

This question—numbers of subtaxa per taxon—is graphically illustrated in Figure 3 in the MP context. We ask, For each ''parental'' lineage at one point in time, how many ''progeny'' lineages (including itself) does it have at the later point in time? An illustration of a bar chart that this generates is given in Figure 4 which is adapted from Nee et al. (1992). We drew lines across Sibley and Ahlquist's (1990) molecular phylogeny of the birds at entirely arbitrary places deep in the tree.

The line in Figure 4 is a fitted geometric distribution, which is what is expected from a pure birth process. In fact, when looking at the entire history of a clade, the geometric distribution is expected for *any* process that has the following property: all lineages have the same probability of speciating (or going extinct, if we are referring to the underlying process generating the clade whose MP we are observing). We do not need to suppose that these probabilities are constant—they can vary arbitrarily (e.g., Nee and May 1994).

From the point of view of hypothesis testing, this is a wonderful fact for the following reason. Conditioning on the number of progeny lineages, the distribution of family sizes under the null hypothesis of lineage equivalence is ''broken stick.'' So, for example, if there are ten parents and 100 progeny, the distribution of family sizes can be determined by randomly breaking a stick 100 units long into ten pieces. This makes for easy statistical analysis of size questions that may be of interest: for example, is there an excess of particularly small taxa (Strathman and Slatkin 1983; Nee et al. 1996)? are some taxa improbably large? This latter question defines adaptive radiations as clades that are anomalously large with respect to a null model of lineage equivalence.

The Passeri (songbirds) and Ciconiiformes (storks, shorebirds, tubenoses, birds of prey, and others.) were identified as statistical outliers and are highlighted in Figure 4. One can speculate as to why these groups radiated so exceptionally: it is surely no coincidence that the Ciconiiform radiation occurred at the time of the breakup of Gondwanaland (Cotgreave and Harvey 1994). Incidentally, these two radiations are, analyzed on their own, an excellent fit to the pure birth process.

There is one problem, however, which arises from the fact that where we draw our lines in Figure 3 is arbitrary. If we just keep moving the parental line all down the phylogeny until we find something interesting, then the multiple-test problem makes statistical signifi-

cances hard to interpret. Also, there are puzzling, unresolved issues about the independence, or lack thereof, of necessarily nested tests. This line-sliding was actually done by Purvis et al. (1995), in a search for primate radiations (they are abundant among Old World monkeys). It seems to me that one can do no more than simply be aware of this problem.

Recent Challenge

It seems appropriate to end by fending off an extraordinary recent challenge to the theoretical framework that has served P and MP so well for so long. Hubbell (2001) rejects the birth and birth-death models on two grounds. The first is that lineages are not all the same with respect to their probabilities of cladogenesis. This must be true, but in itself in no way does it undermine the utility of these models as *null* models for statistical purposes. Superficially more serious is the claim that the data themselves are clearly incompatible with a null model that predicts geometric distributions of subtaxa per taxon, such as we saw for Sibley and Ahlquist's bird phylogeny in Figure 4, and he remarkably supports this claim with data from Sibley and Ahlquist's bird phylogeny (Hubbell 2001: Fig. 8.4).

Hubbell's claim is based on a misapprehension. The representation of the data in Figure 4 is appropriate for geometrically distributed data with a small mean value, so that the probabilities of one, two, three, and so on are high. If, on the other hand, the mean is large, then the probability of any *particular* number becomes negligible, and a frequency histogram approach, such as we will see below, becomes the sensible one. The data that Hubbell investigates are the numbers of species per bird family and he plots a frequency histogram of the logarithms of the family sizes. This is appropriate. He observes an interior mode in this distribution and mistakenly asserts that this is incompatible with a geometric distribution: ''In all cases [of geometric distributions with different parameters] the most frequent category represents lineages with only one living descendent (themselves) and the frequency of lineages with a higher number of descendents falls off exponentially. The longer the time period sampled, the larger the number of possible descendent lineages and the frequency distribution becomes flatter; but the highest frequency category remains the singleton category'' (Hubbell 2001: p. 240).

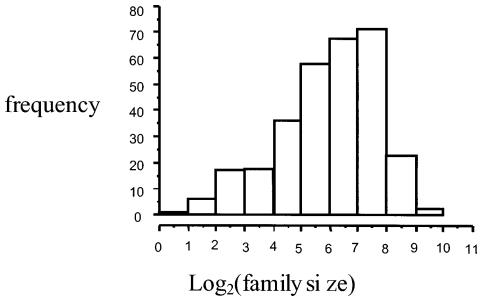

FIGURE 5. Frequency histogram of family sizes. For 300 parental lineages, I drew their number of progeny lineages from a geometric distribution with a mean of 100. After a $\log_2$ transformation of family size (base 2 is traditional in ecology), the figure is a frequency histogram showing how many parental lineages had the number of progeny lineages in each of the bins. In terms of the untransformed family sizes, for each integer n along the x axis, each bin spans family sizes from 2^n to 2^{n+1}.

Certainly, geometrically distributed data, when visualized as in Figure 4, do not exhibit an interior mode. But they do when visualized as a frequency histogram. Figure 5 shows the frequency distribution of the logarithm of the numbers of progeny lineages that have been drawn from a geometric distribution with a large mean—100: its interior mode is obvious. Hubbell's observation of an interior mode in his analysis is in no way inconsistent with the modeling framework discussed here.

Conclusion

Since 1924, simple models have been used to help guide our understanding of the diversification of life. Because paleontology and molecular phylogenetics deal with such different data, it is unsurprising that they have developed separately and that the extensive conceptual overlap that exists between them has gone largely unnoticed. I hope this paper will contribute to forging a bridge between the two intellectual traditions, which are, after all, really interested in precisely the same thing: life on Earth.

Acknowledgments

I am grateful to J. Alroy, M. Foote, H. Sims, and P. Wagner for their helpful comments.

Literature Cited

Bailey, N. T. J. 1964. The elements of stochastic processes with applications to the natural sciences. Wiley, New York.

Baldwin, B. G., and M. J. Sanderson. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proceedings of the National Academy of Sciences USA 95:9402–9406.

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

Cotgreave, P., and P. H. Harvey. 1994. Associations among biogeography, phylogeny and bird species diversity. Biodiversity Letters 2:46–55.

Cox, D. R., and P. A. A. Lewis. 1966. The statistical analysis of series of events. Methuen, London.

Feller, W. 1957. An introduction to probability theory and its applications, Vol. 1. Wiley, New York.

Foote, M. 1988. Survivorship analysis of Cambrian and Ordovician trilobites. Paleobiology 14:258–271.

———. 2001. Evolutionary rates and the age distributions of living and extinct taxa. Pp. 245–295 *in* J. B. C. Jackson, S. Lidgard, and F. K. McKinney, eds. Evolutionary patterns: growth, form and tempo in the fossil record. University of Chicago Press, Chicago.

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

Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46:627–640.

Hillis, D. M., C. Moritz, and B. K. Mable. 1996. Molecular systematics. Sinauer, Sunderland, Mass.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, NJ.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. *In* D. Futuyma and J. Antonovics, eds. Oxford Surveys in Evolutionary Biology 8:1–49.Oxford University Press, Oxford.

Kendall, D. G. 1948. On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. Biometrika 35:6–15.

———. 1949. Stochastic processes and population growth. Journal of the Royal Statistical Society B 11:230–264.

Kitchell, J. K., and N. MacLeod. 1988. Macroevolutionary interpretations of symmetry and synchroneity in the fossil record. Science 240:1190–1193.

Losos, J. B., and F. R. Adler. 1995. Stumped by trees? A generalized null model for patterns of organismal diversity. American Naturalist 145:329–342.

Moran, P. A. P. 1958. Random processes in genetics. Proceedings of the Cambridge Philosophical Society 54:60–71.

Nee, S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661–668.

Nee, S., and R. M. May. 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London B 344:305–311.

Nee, S., A. O. Mooers, and P. H. Harvey. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proceedings of the National Academy of Sciences USA 89:8322–8326.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1995. Estimating extinction from molecular phylogenies. Pp. 164–182 *in* J. L. Lawton and R. M. May, eds. Extinction rates. Oxford University Press, Oxford.

Nee, S., T. Barraclough, and P. H. Harvey. 1996. Temporal changes in biodiversity: detecting patterns and identifying causes. Pp. 230–252 *in* K. J. Gaston, ed. Biodiversity: an introduction. Blackwell Science, Oxford.

Paradis, E. 1997. Assessing temporal variation in diversification rates from phylogenies: estimation and hypothesis testing. Proceedings of the Royal Society of London B 264:1141–1147.

Purvis, A., S. Nee, and P. H. Harvey. 1995. Macroevolutionary inferences from primate phylogeny. Proceedings of the Royal Society of London B 260:329–333.

Raup, D. M. 1975. Taxonomic survivorship curves and Van Valen's law. Paleobiology 1:82–96.

———. 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. Journal of Geology 81:525–542.

Sepkoski, J. J., Jr. 1978. A kinetic model of Phanerozoic taxonomic diversity. I. Analysis of marine orders. Paleobiology 4:223–251.

Sepkoski, J. J., Jr., and D. C. Kendrick. 1993. Numerical experiments with model monophyletic and paraphyletic taxa. Paleobiology 19:168–184.

Sibley, C. G., and J. E. Ahlquist. 1990. Phylogeny and classification of birds. Yale University Press, New Haven, Conn.

Stanley, S. M. 1975. A theory of evolution above the species level. Proceedings of the National Academy of Sciences USA 72: 646–650.

Stanley, S. M., and W. A. Newman. 1980. Competitive exclusion in evolutionary time: the case of the acorn barnacles. Paleobiology 6:173–183.

Strathman, R. R., and M. Slatkin. 1983. The improbability of animal phyla with few species. Paleobiology 9:97–106.

Uhen, M. D. 1996. An evaluation of clade-shape statistics using simulations and extinct families of mammals. Paleobiology 22:8–22.

Yule, G. U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, FRS. Philosophical Transactions of the Royal Society of London B 213:21–87.