Evolution, 46(6), 1992, pp. 1818–1826

PATTERNS IN TREE BALANCE AMONG CLADISTIC, PHENETIC, AND RANDOMLY GENERATED PHYLOGENETIC TREES

STEPHEN B. HEARD
Department of Biology, Leidy Labs, University of Pennsylvania, Philadelphia, PA 19104-6018 USA

Abstract.—I examine patterns in tree balance for a sample of 208 cladograms and phenograms from the recent literature. I provide an expression for expected imbalance under a simple, uniform-rate random speciation model, and I estimate variances by simulation for the same model. Imbalance decreases with tree size (number of included taxa) in both theoretical and literature trees. In contrast to previous suggestions, I find cladistic trees to be no more imbalanced than phenetic trees when confounding variables are appropriately controlled. The degree of imbalance found in literature trees is inconsistent with the uniform-rate speciation model; this is most likely a result of variability in speciation and extinction rates among real lineages. The existence of such variation is a necessary (but not sufficient) condition for the operation of the macroevolutionary processes of species sorting and species selection.

Key words.—Balance, extinction rates, phylogenies, speciation rates, species sorting, tree topology.

Received July 12, 1991. Accepted April 10, 1992.

Phylogenetic trees, and their (intended) close relatives cladograms and phenograms, vary in their balance—that is, the degree to which branch points define subgroups of equal size (Fig. 1). Balance is an interesting attribute of phylogenetic trees because patterns in balance can reveal important features of the underlying evolutionary process. In particular, balance is influenced by the degree of variability of speciation and extinction rates among lineages; the more variable either rate, the less balanced (on average) the resulting trees. Variability in these rates provides the raw material for species sorting (sensu Eldredge, 1989), which has been suggested (e.g., Vrba and Eldredge, 1984) to reflect important macroevolutionary processes.

Patterns in balance of *estimated* phylogenetic trees (the cladograms and phenograms produced by systematists) might, however, also reflect choices in definition and inclusion of taxa to be analyzed or properties of the estimation algorithms. For instance, it has been suggested (e.g., Colless, 1982; Shao and Sokal, 1990) that cladistic trees tend to be more imbalanced than phenetic trees (but see Savage, 1983). Such possible methodological influences must be ruled out or controlled before patterns in balance can be related to evolutionary processes, or before balance can be compared between clades.

Finally, the degree of balance in any particular clade is partly a result of stochastic events in its evolution. We need to disentangle the effects of evolutionary process, methodology, and stochasticity in shaping tree balance. Statistical analysis of large sets of estimated trees can achieve this end by identifying patterns in tree balance and partitioning variance in balance among sources.

I examined patterns in balance of 208 estimated trees, both cladistic and phenetic, taken from the recent literature. I applied an index (defined below) which measures imbalance, ranging from 0 for a perfectly balanced tree (Fig. 1a) to 1 for a perfectly imbalanced one (Fig. 1b). I used this data set to explore three aspects of the determinants of balance.

First, I examined some possible effects of methodology on balance. I tested the suggestion of Colless (1982) and Shao and Sokal (1990) that cladistic techniques yield more imbalanced trees than phenetic techniques. If this is true, one method must have a systematic bias in tree estimation, which would be unfortunate as each method is often (although not always) intended to reconstruct true evolutionary trees. I also examined two other features of systematic analyses: type of data (molecular or morphological) and taxonomic rank of the analyzed taxa. Differences in balance among taxonomic ranks have been addressed by

Dial and Marzluff (1989) and Anderson (1974), using classifications rather than phylogenies, and by Guyer and Slowinski (1991).

Second, I examined trends in balance with tree size (number of included taxa). Expected frequencies of tree topologies vary with tree size under several theoretical models (Savage, 1983) for combinatorical reasons (Page, 1991). Therefore, it is very likely that size will also affect the balance of samples of estimated trees.

Third, having controlled or rejected methodological influences, I compared the balance of estimated trees to that predicted by a simple model of random evolution with constant rates of speciation and extinction. Deviation of balance from this model can result from variability in speciation or extinction rates among real lineages, as suggested by Raup et al. (1973). I discuss some implications of patterns in balance for macroevolutionary models.

Materials and Methods

1. Index of Imbalance

I used a corrected version of Colless' (1982) index to assess imbalance. I used a calculated statistic rather than analyzing frequency distributions (e.g., Guyer and Slowinski, 1991) because it facilitates comparisons and because for large trees the number of possible topologies becomes extremely unwieldy. The index is computed as follows: for every interior node in a tree of n taxa, count the number of terminal taxa subtended by the right hand branch (T_R) and the number subtended by the left hand branch (T_L). Then calculate:

$$Im = \frac{\sum_{\text{all interior nodes}} |T_R - T_L|}{(n-1)(n-2)/2} \quad (1)$$

which ranges from 0 (perfect balance; Fig. 1a) to 1 (complete imbalance; Fig. 1b). A score of 0 may be attained only by a tree with n equal to a power of 2; otherwise, perfect balance is not possible. The normalizing denominator is corrected from Colless (1982), who mistakenly used $[(n)(n - 3) + 1]/2$.

Shao and Sokal (1990) discuss several indices which measure imbalance in slightly

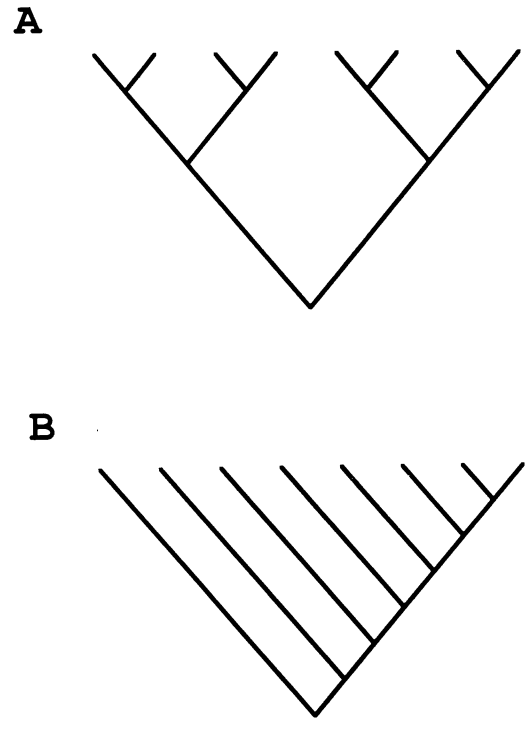

FIG. 1. Completely balanced (A) and imbalanced (B) 8-taxon phylogenetic trees.

different ways. I have used *Im* because it is simple and intuitive; it is highly correlated with other indices. Three points are germane to the choice of an index.

First, *Im* cannot be applied to trees with polytomies, whereas some indices can be. However, polytomies probably represent incomplete knowledge ("unresolved nodes") far more often than they do true multifurcative evolution, and little is to be gained from the study of the topology of such trees (although particular well-resolved clades within such trees might be of interest). Unless the polytomies are thought to indicate real multifurcations, a tree with such nodes should be disregarded in studies of balance.

Second, some of Shao and Sokal's (1990) indices are normalized to range from 0 to 1 *for every tree size.* However, normalization to 0 is unsatisfactory because a tree whose size is not a power of 2 can never be perfectly balanced (recall Fig. 1a). *Im* as I define it does not have this disadvantage; the other indices could be improved by normalizing to 1 but not to 0.

Finally, Shao and Sokal (1990) recom-

mend the use of normalized indices [their $I_C(3)$, $I_S(3)$, $B_1(3)$, and $B_2(3)$] to allow comparison of trees with different numbers of taxa. As I establish below, tree size is indeed an important determinant of balance. However, none of their indices is actually size-independent: expected values (under the random speciation model) of $I_C(3)$ and $I_S(3)$ decline with tree size, while $B_2(3)$ increases and $B_1(3)$ changes nonmonotonically. Values for estimated trees behave similarly. Tree size must be taken into account with any index, at least for small trees or trees of markedly different sizes.

2. Balance under a Simple Random Evolution Model

Expected distributions of the balance index were obtained for a null hypothesis of random, constant rate speciation. This is equivalent to the "Markov null model" of Simberloff et al. (1981). Results apply equally to a somewhat more elaborate model which incorporates random extinction.

The expected value of *Im* for samples of n-trees is given by

$$\mathrm{E}(Im) = \begin{cases} \dfrac{2n}{(n-1)(n-2)} \displaystyle\sum_{j=2}^{n/2} \frac{1}{j} & (n \text{ even}) \\ \dfrac{2n}{(n-1)(n-2)} \left[\dfrac{1}{n} + \displaystyle\sum_{j=2}^{(n-1)/2} \frac{1}{j} \right] & (n \text{ odd}) \end{cases} \quad (2)$$

A derivation of equation (2) is provided in the Appendix. However, to find the variance of *Im* analytically (needed for statistical testing) would be exceedingly complex. Frequency distributions could also, in theory, be worked out using the methods of Page (1991), but as the number of possible topologies increases very rapidly with tree size (Simberloff et al., 1981), this would be prohibitively time-consuming for trees larger than five or six taxa.

A simulation method proved to be the most practical for obtaining the variances of balance distributions, and for drawing subsamples to compare with samples of observed trees. A computer program (written in BASIC) was therefore used to generate balance distributions. The algorithm began with a pair of taxa, and at each iteration chose one at random to split (speciate), giving a tree with one more taxon than in the previous iteration. Each run produced one phylogenetic tree of the desired size, with a particular topology and set of divergence times (equivalent to a dendrogram in Page, 1991). Ten thousand trees were generated for each tree size from 4 to 14, and *Im* calculated and recorded for each. Mean values for the simulation results matched the theoretical values (equation 2) closely. Although for the smallest tree sizes distributions could have been worked out by hand, I chose to use the simulated distributions for the sake of consistency.

At least two other models for expected topology frequencies have been discussed: the equiprobable-trees model (Simberloff et al., 1981) and the proportional-to-distinguishable-types model (Rosen, 1978; Simberloff et al., 1981). However, these are based only on assumed uniform distributions of tree shapes; they might be realistic only if systematists did no better at tree reconstruction than random choices among all possible trees. Only the random speciation model is underlain by any plausible evolutionary process (Page, 1991; Savage 1983). The failure of the equiprobable and distinguishable-types models to fit evolutionary data (Savage, 1983; but see Guyer and Slowinski, 1991, for 5-trees) is therefore unsurprising, and I discuss these models no further.

3. Data Compilation

Recent issues of 12 journals were surveyed for published trees (Table 1). All trees that met the following conditions were included: a) Between 4 and 14 taxa included. Sample sizes for larger trees would have been too small for useful analysis; only one topology is possible for smaller trees. b) No polytomies. As discussed above, *Im* is applicable only to fully resolved trees, and little would be gained from the study of the topology of unresolved trees. c) Analyzed taxa of consistent rank. Trees including taxa of very different rank (e.g., "Arthropods" and "*Mus musculus*") were omitted. In a very few cases, trees with most taxa being species (or genera) but a few species (or genera) further broken down were included af-

ter lumping the taxa of lower rank. d) Method of analysis (cladistic/phenetic) and type of data (molecular/morphological) reported (only a few trees lacked this information). I did not discriminate among methods within the phenetic and cladistic classes, although there may be some differences (e.g., UPGMA versus single-linkage phenograms; Rohlf, 1982). I did not distinguish between strongly and weakly supported trees, nor did I require that trees included all members of the presumed monophyletic group. The latter point was addressed by Guyer and Slowinski (1991), who found that random omission of taxa from "complete" trees did not affect conclusions about tree balance.

For each selected tree, I recorded size, the imbalance index *Im,* the type of analysis (cladistic or phenetic), the type of data (molecular or morphological), and the rank of the analyzed taxa (populations/subspecies, species, or higher taxa). If the tree included outgroups, the number of taxa and *Im* were recorded both with and without the outgroup(s). The values without outgroups were used in all analyses (except as noted), because outgroups (by definition) are not part of the groups of primary interest in systematic analyses.

In a few cases, two trees based on different analysis types or data types were reported for a single set of taxa. In these cases, if an analysis drew a contrast such that one tree was in each group, both were retained; but otherwise, one tree was selected at random to avoid pseudoreplication. A full list of trees used is available on request.

4. Comparisons and Statistical Testing

A number of tests were used to examine patterns in *Im,* contrasting different sets of estimated trees and comparing estimated to randomly generated trees. Expected distributions of *Im* are discrete and highly nonnormal; therefore, randomization and Monte Carlo methods [using microcomputer programs written in BASIC and SAS Version 6.03 (SAS, 1988)] were applied. These methods eliminate the need to meet the continuous distribution and normality assumptions of standard statistical testing (Manley, 1991). Because not all the analyses use standard packaged procedures, I describe the major ones in some detail here.

TABLE 1. Journals surveyed for cladograms and phenograms.

Journal	Volumes[a]
Annals of the Entomological Society of America	78–83
Biological Journal of the Linnean Society	24–36, 39–41[b]
Botanical Journal of the Linnean Society	90–98, 102–104[b]
Canadian Journal of Botany	63–68
Canadian Journal of Zoology	63–68
Cladistics	1–6
Evolution	39–44
Molecular Biology and Evolution	3–7
Plant Systematics and Evolution	149–173
Systematic Botany	10–15
Systematic Zoology	34–39
Zoological Journal of the Linnean Society	84–95, 98–100[b]

[a] Covering 1985 to 1990 inclusive.
[b] Some volumes not surveyed due to incomplete library holdings.

a) Sources of variance in *Im* among estimated trees. I began by running two 4-way ANOVAs (SAS, PROC GLM) on the compiled data: one using standard parametric methods, and the other with all *Im* values replaced with their ranks. The discrete, nonnormal distributions of *Im* violate the assumptions of the parametric test, and the fact that the distributions of *Im* differ across tree sizes violates the assumptions of both tests. However, they were still useful as a first pass at the data. In particular, since the type of data (morphological versus molecular) explained almost no variance ($F < 0.25$) in either test, this variable was dropped from subsequent analyses.

To examine the importance of analysis type and taxon rank, I performed separate analyses for each tree size (4 to 14), except that I was forced to pool some larger size trees (10 with 11, 12 and 13 with 14) to attain useful sample sizes. This procedure avoids difficulty with the dissimilarity of distributions of *Im* for small trees; for larger trees *Im* distribution changes relatively little in mean (Fig. 2) or shape. For each size, I ran a 2-way ANOVA (SAS PROC GLM), evaluating significance by randomization (as

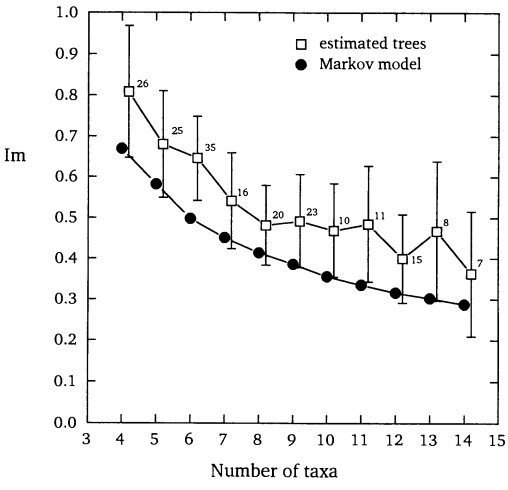

FIG. 2. Mean imbalance for 195 estimated trees and expected imbalance for the random, uniform-rate specication model. Bars for estimated trees are two standard errors (jackknifed estimates); due to severe non-normality, for the smaller sample sizes two standard errors are only a very approximate 95% C.I. Numbers with data points are sample sizes. Expected values for the random model are from equation 2; symbols cover two standard errors (bootstrapped, from simulations, 10,000 trees for each size).

detailed in Manley, 1991). I combined the tests for different tree sizes, which are independent under the null hypothesis of no effects, using Fisher's method (Manley, 1985, his case 2). This procedure yields a powerful, single chi-square test, while still controlling effectively for tree size.

Finally, to verify the effect of size in estimated trees, I used a regression analysis, pooling across analysis type and taxon rank (nonsignificant in the test above). Although the *Im* values themselves are discretely and non-normally distributed, the mean values for each tree size should be much less so (central limit theorem). Therefore, I regressed the means against tree size, weighting by sample sizes (Kleinbaum et al., 1988).

b) Comparison of estimated trees with the random, uniform-rate speciation model. Mean *Im* values for estimated trees were tabulated by tree size, pooling across data type, analysis type, and taxon rank. Because of the severe non-normality of expected balance distributions, each tree size sample was compared to the distributions for the random model using a Monte Carlo procedure. Ten thousand samples (of the appropriate number of trees) were drawn from the simulation *Im* distribution for that tree size, and the fraction equal to or exceeding the observed values was recorded. These independent *P*-values were combined using Fisher's method to give a single overall test. This procedure treats the sample of 10,000 simulations as if it were the population; the error introduced by this approximation is very small.

RESULTS AND DISCUSSION

1. Imbalance of Randomly Generated Trees

Expected values of *Im* for trees generated under the model of random speciation (and extinction) decline markedly with tree size, even though *Im* is normalized to range from 0 to 1 (Eq. 2; Fig. 2). This has the important implication that comparisons of imbalance must always be controlled for tree size. The importance of tree size for imbalance echoes its importance in determining phenogram estimation accuracy (Astolfi et al., 1981), and suggests that tree size might influence other tree properties. I control for tree size in all analyses reported here, except as noted ("naive" test, below).

2. Imbalance of Estimated Trees

In the 4-way (data type, analysis type, taxon rank, tree size) ANOVAs, data type explained almost no variance (parametric $F = 0.08$, $P \approx 0.8$; ranked $F = 0.21$, $P \approx 0.6$). The unimportance of this factor is so clear that more robust (and complicated) analyses were considered unnecessary. Data type was omitted from subsequent analyses.

Analysis type and taxon rank had no significant effects on balance in the combined 2–way ANOVAs (analysis type: $X^2 = 20.59$, $df = 16$, $P = 0.805$; taxon rank: $X^2 = 20.56$, $df = 16$, $P = 0.804$). Further analyses pooled across both factors. Mean imbalance values are plotted in Figure 2.

The unimportance of analysis type contrasts with the suggestions of Colless (1982) and Shao and Sokal (1990). In fact, a "naive" *t*-test, which deliberately neither controlled tree size nor removed outgroups from cladograms, found cladograms less balanced (although the difference was of marginal significance; randomization $P = 0.059$). This can be accounted for on the grounds that cladistic trees were smaller (Wilcoxon

test, normal approximation, $Z = 2.82$, $P = 0.0048$; and small trees are more imbalanced), and trees with outgroups (always cladograms) were more imbalanced (43 more, 7 less, 7 equal; sign test $P < 0.0001$). Clearly, although artificial data sets can be constructed for which the methods disagree markedly on topology (E. Theriot, pers. comm.), comparable samples of real cladistic and phenetic trees are indistinguishable in terms of balance; the previous suggestions appear to have been ill-founded.

The unimportance of taxon rank is consistent with the results of Dial and Marzluff (1989) for classifications. On the other hand, Guyer and Slowinski (1991) found that trees of genera were more balanced than those of species. This lack of agreement is puzzling, and the topic is worthy of further examination.

Regression analysis confirmed the importance of tree size. When data were pooled across taxon rank and analysis type, tree size explained 82% of the variation in mean tree imbalance (mean imbalance = $0.878 - 0.039 \times$ (tree size), $R^2 = 0.82$; $P < 0.0001$). This result reinforces the conclusion that tree size must be considered in analyzing balance.

3. Estimated Trees and the Random Speciation Model

Estimated trees were much less balanced than expected from the random speciation model (combined test: $X^2 = 108.06$, $df = 22$, $P < 0.0001$; see Fig. 2). Guyer and Slowinski (1991) and Raup et al. (1973) reported similar discrepancies. Savage (1983) found significant differences (between topology frequencies of estimated and random-speciation trees) for 5-member trees, but not for 4-, 6-, or 7-member trees. He did not comment on any differences in mean imbalance, and because his sets of trees of different sizes were not independent (some small trees were subsets of larger trees) his tests for different sizes cannot be combined.

The equivalent phenomenon in classifications is also familiar, although rarely explicitly tested: one or a few subtaxa often account for much of the diversity of a larger taxon. For instance, most mammals are rodents, most birds are passerines, and most insects are beetles (e.g., Anderson, 1974; Willis and Yule, 1922; Dial and Marzluff, 1989).

A tendency for speciation or extinction rates to vary randomly among lineages would produce more imbalanced trees. Grant (1963) and Stanley (1979) have pointed out that *non*-random variation of rates would also produce imbalanced trees. Although many studies have implied or demonstrated variable speciation or extinction rates (Bush et al., 1977; Eldredge, 1989; Levin and Wilson, 1976; Raup et al., 1973; Simpson, 1953; Stanley et al., 1981; papers in Eldredge and Stanley, 1984), the degree, causes, and implications of such variability are contentious (e.g., Eldredge, 1984; Stanley 1984).

Variability of speciation or extinction rates provides the necessary raw material for the operation of species sorting (Eldredge, 1989). That patterns in balance of estimated trees are inconsistent with the uniform-rate model corroborates (from a novel angle) the existence of such variability, although it need not necessarily imply an important role of species sorting in macroevolution. Variable rates have also been held necessary to account for observed radiations of some groups (e.g., Bivalvia since the Triassic; Stanley et al., 1981).

The balance predictions of more sophisticated speciation models, incorporating specific patterns of variable rates, would be worthy of attention. When corrected for tree size, comparisons of balance using *Im* on cladograms, phenograms, and simulated trees provide a direct, powerful approach to the study of patterns in branching evolution. Studies comparing balance among different theoretical models, major clades or ecological types could prove rewarding.

Acknowledgments

I thank E. Cooch and V. Apanius for statistical suggestions. W. Ewens outlined the derivation in the Appendix. M. Donoghue, J. Huelsenbeck, R. Page, N. Shubin, and two anonymous reviewers made valuable comments on the manuscript. While conducting this research I was supported by a Natural Sciences and Engineering Research Council (Canada) "1967" scholarship.

Literature Cited

Anderson, S. 1974. Patterns of faunal evolution. Q. Rev. Biol. 49:311–332.

Astolfi, P., K. K. Kidd, and L. L. Cavalli-Sforza.

1981. A comparison of methods for reconstructing evolutionary trees. Syst. Zool. 30:156–169.

BUSH, G. L., S. M. CASE, A. C. WILSON, AND J. L. PATTON. 1977. Rapid speciation and chromosomal evolution in mammals. Proc. Natl. Acad. Sci. USA 74:3942–3946.

COLLESS, D. H. 1982. Phylogenetics: The theory and practice of phylogenetic systematics II (book review). Syst. Zool. 31:100–104.

DIAL, K. P., AND J. M. MARZLUFF. 1989. Nonrandom diversification within taxonomic assemblages. Syst. Zool. 38:26–37.

ELDREDGE, N. 1984. Simpson's inverse: bradytely and the phenomenon of living fossils, pp. 272–277. *In* N. Eldredge and S. M. Stanley (eds.), Living Fossils. Springer-Verlag, N.Y., USA.

———. 1989. Macroevolutionary Dynamics: Species, Niches, and Adaptive Peaks. McGraw-Hill, N.Y., USA.

ELDREDGE, N., AND S. M. STANLEY (eds.) 1984. Living Fossils. Springer-Verlag, N.Y., USA.

FELLER, W. 1968. An Introduction to Probability Theory and Its Applications. Vol. 1. Third ed. John Wiley and Sons, N.Y., USA.

GRANT, V. 1963. The Origin of Adaptations. Columbia University Press, N.Y., USA.

GUYER, C., AND J. B. SLOWINSKI. 1991. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. Evolution 45:340–350.

KLEINBAUM, D. G., L. L. KUPPER, AND K. E. MULLER. 1988. Applied Regression Analysis and Other Multivariable Methods. Second ed. P. W. S. Kent, Boston, MA USA.

LEVIN, D. A., AND A. C. WILSON. 1976. Rates of evolution in seed plants: Net increase in diversity of chromosome numbers and species numbers through time. Proc. Natl. Acad. Sci. USA 73:2086–2090.

MANLEY, B. F. J. 1985. The Statistics of Natural Selection on Animal Populations. Chapman and Hall, London, UK.

———. 1991. Randomization and Monte Carlo Methods in Biology. Chapman and Hall, London, UK.

PAGE, R. D. 1991. Random dendrograms and null hypotheses in cladistic biogeography. Syst. Zool. 40:54–62.

RAUP, D. M., S. J. GOULD, T. J. M. SCHOFF, AND D. S. SIMBERLOFF. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81: 525–542.

ROHLF, F. J. 1982. Consensus indices for comparing classifications. Math. Biosci. 59:131–144.

ROSEN, D. E. 1978. Vicariant patterns and historical explanation in biogeography. Syst. Zool. 27:159–188.

SAS. 1988. SAS/STAT User's Guide, Release 6.03 edition. SAS Institute Inc., Cary, NC USA.

SAVAGE, H. M. 1983. The shape of evolution: systematic tree topology. Biol. J. Linn. Soc. 20:225–244.

SHAO, K., AND R. R. SOKAL. 1990. Tree balance. Syst. Zool. 39:266–276.

SIMBERLOFF, D., K. L. HECK, E. D. MCCOY, AND E. F. CONNOR. 1981. There have been no statistical tests of cladistic biogeographical hypotheses, pp. 40–63. *In* G. Nelson and D. E. Rosen (eds.), Vicariance Biogeography: A Critique. Columbia University Press, N.Y., USA.

SIMPSON, G. G. 1953. The Major Features of Evolution. Columbia University Press, N.Y., USA.

STANLEY, S. M. 1979. Macroevolution: Pattern and Process. W. H. Freeman, San Francisco, CA USA.

———. 1984. Does bradytely exist? pp. 278–280. *In* N. Eldredge and S. M. Stanley (eds.), Living Fossils. Springer-Verlag, N.Y., USA.

STANLEY, S. M., P. W. SIGNOR III, S. LIDGARD, AND A. F. KARR. 1981. Natural clades differ from "random" clades: simulations and analyses. Paleobiology 7:115–127.

VRBA, E. S., AND N. ELDREDGE. 1984. Individuals, hierarchies, and processes: towards a more complete evolutionary theory. Paleobiology 10:146–171.

WILLIS, J. C., AND G. U. YULE. 1922. Some statistics of evolution and geographical distribution in plants and animals, and their significance. Nature 109: 177–179.

Corresponding Editor: M. Donoghue

APPENDIX

Derivation for Expected Imbalance under Random Speciation

Recall that the imbalance index is:

$$Im = \frac{\sum_{\text{all interior nodes}} |T_R - T_L|}{(n-1)(n-2)/2} \qquad \text{(A1)}$$

where n is the number of terminal taxa in the tree, and T_R and T_L are the numbers of terminal taxa arising from the right and left branches of each interior node.

The derivation of the mean value of Im, assuming random branching, is in three parts. First, we find the expected number of j-nodes (nodes with exactly j descendent taxa) in an n-tree. Second, we determine the expected value of $|T_R - T_L|$ for a j-node. Finally, we combine these expected values in equation A1. The logic is easiest to follow when related to a tree as it grows (evolves) rather than to the static end product. Recall that in a uniform-rate random speciation process, at any given time any of the extant taxa in a tree is equally likely to be involved in the next speciation event.

Part 1. Number of j-nodes in an n-tree.

Consider a tree of n terminal taxa. The earliest speciation event split the common ancestor into two descendents. This is the root of the tree; label it node 1. The subsequent proliferation of the clade involved $(n - 1)$ further bifurcations; label the corresponding nodes as node 2, 3, . . . $(n - 1)$.

Now consider node i (corresponding to the i^{th} speciation event). What is the probability that this node will have j descendent terminal taxa when the tree has grown to size n? Node i, after the speciation event, has two immediate descendent taxa; the rest of the tree has $(i - 1)$ taxa. The next speciation event will occur at random among the $(i + 1)$ extant taxa. This situation is equivalent to repeatedly drawing balls from an urn which initially contains two black balls and $(i - 1)$ red

balls, if at each drawing the chosen ball is replaced and another of the same color is added (a Polya urn; Feller, 1968). After $(n - i - 1)$ such draws, there will be n balls; the probability we want is that of drawing $(j - 2)$ black and $(n - i - j + 1)$ red balls, leaving us with a total of j blacks and $(n - j)$ reds (i.e., j descendents from our node i). This probability is given (Feller, 1968) by

$$P(j \mid i) = \begin{cases} 1, & i = 1, j = n \\ (j-1)\cdot\dfrac{\binom{n-j-1}{n-i-j+1}}{\binom{n-1}{n-i-1}}, & \text{otherwise.} \end{cases}$$

So, the expected number of j-nodes in an n-tree is

$$\text{E(\# } j\text{-nodes)} = \begin{cases} 1, & j = n \\ (j-1)\cdot\displaystyle\sum_{i=2}^{n-1}\frac{\binom{n-j-1}{n-i-j+1}}{\binom{n-1}{n-i-1}}, & j < n \end{cases}$$

Working with the case $j < n$,

$$\begin{aligned} &\text{E(\# } j\text{-nodes)} \\ &= (j-1)\cdot\sum_{i=2}^{n-j+1}\frac{\binom{n-j-1}{n-i-j+1}}{\binom{n-1}{n-i-1}} \\ &\quad + (j-1)\cdot\sum_{i=n-j+2}^{n-1}\frac{\binom{n-j-1}{n-i-j+1}}{\binom{n-1}{n-i-1}} \\ &= (j-1)\cdot\sum_{i=2}^{n-j+1}\left[\frac{(n-j-1)!(n-i-1)!i!}{(n-1)!(i-2)!(n-i-j+1)!}\right] + 0 \\ &= \frac{2n}{j(j+1)}\cdot\sum_{i=2}^{n-j+1}\left[\frac{(j+1)!(n-j-1)!}{2(j-1)!(i-2)!(n-i-j+1)!}\right. \\ &\qquad \left.\cdot\frac{(n-i)!i!}{n!}\cdot\frac{j-1}{n-1}\right] \\ &= \frac{2n}{j(j+1)}\cdot\sum_{i=2}^{n-j+1}\left[\frac{\binom{j+1}{2}\binom{n-j-1}{i-2}}{\binom{n}{i}}\cdot\frac{j-1}{n-i}\right] \end{aligned} \tag{A2}$$

But the summation in equation A2 is equal to 1, because it is a probability distribution (in particular, for the probability of needing exactly $(i + 1)$ draws to pick three red balls (without replacement) from an urn containing $(j + 1)$ red and $(n - j - 1)$ black balls). Therefore we have simply

$$\text{E(\# } j\text{-nodes)} = \begin{cases} 1, & j = n \\ \dfrac{2n}{j(j+1)}, & j < n \end{cases}$$

Part 2. Expected value of $|T_R - T_L|$ for a j-node.

Any j-node in the tree began as a speciation event yielding two taxa; each subsequent speciation event then split one descendent taxon at random into two. This is equivalent to a Polya urn beginning with one black and one red ball, with drawings continuing until there are j balls in the urn (Feller, 1968). The distribution of T_R is therefore uniform on $[1, 2, \ldots (j - 1)]$, and T_L is just $(j - T_R)$, so the expected value is:

$$\mathrm{E}_j\,|T_R - T_L| = \begin{cases} \dfrac{(j-1)}{2}, & j \text{ odd} \\ \dfrac{j(j-2)}{2(j-1)}, & j \text{ even} \end{cases} \tag{A3}$$

Part 3. Expected value of Im.

$$\begin{aligned} &\mathrm{E}(Im) \\ &= \mathrm{E}\left[\frac{\sum_{\text{all interior nodes}}|T_R - T_L|}{(n-1)(n-2)/2}\right] \\ &= \frac{2}{(n-1)(n-2)}\cdot\sum_{j=2}^{n}[\text{E(\# } j\text{-nodes)}]\cdot[\mathrm{E}_j\,|T_R - T_L|] \\ &= \frac{2n}{(n-1)(n-2)}\cdot\left[\sum_{j=2}^{n-1}\frac{2}{j(j+1)}\cdot\mathrm{E}_j\,|T_R - T_L|\right. \\ &\qquad \left. + \frac{1}{n}\cdot\mathrm{E}_n\,|T_R - T_L|\right] \end{aligned}$$

Substituting appropriately from equation A3, we have:

Case 1: for n even.

$$\begin{aligned} &\mathrm{E}(Im) \\ &= \frac{2n}{(n-1)(n-2)}\cdot\left[\sum_{j=3,5}^{n-1}\frac{j-1}{j(j+1)}\right. \\ &\qquad + \sum_{j=2,4}^{n-2}\frac{j-2}{(j-1)(j+1)} \\ &\qquad \left. + \frac{n-2}{2(n-1)}\right] \\ &= \frac{2n}{(n-1)(n-2)}\cdot\left[2\cdot\sum_{j=3,5}^{n-1}\frac{1}{j+1} - \sum_{j=3,5}^{n-1}\frac{1}{j}\right. \\ &\qquad \left. + \frac{3}{2}\cdot\sum_{j=2,4}^{n-2}\frac{1}{j+1}\right. \end{aligned}$$

$$- \frac{1}{2} \cdot \sum_{j=2,4}^{n-2} \frac{1}{j-1} + \frac{n-2}{2(n-1)} \Bigg]$$

$$= \frac{2n}{(n-1)(n-2)} \cdot \Bigg[2 \cdot \sum_{j=3,5}^{n-1} \frac{1}{j+1} + \left[\frac{1}{2(n-1)} - \frac{1}{2} + \frac{n-2}{2(n-1)} \right] \Bigg]$$

$$= \frac{2n}{(n-1)(n-2)} \cdot \left[2 \cdot \sum_{j=3,5}^{n-1} \frac{1}{j+1} + 0 \right]$$

$$= \frac{2n}{(n-1)(n-2)} \cdot \sum_{i=2}^{n/2} \frac{1}{j}$$

Case 2: for n odd.

E(Im)

$$= \frac{2n}{(n-1)(n-2)} \cdot \left[\sum_{j=3,5}^{n-2} \frac{j-1}{j(j+1)} + \sum_{j=2,4}^{n-1} \frac{j-2}{(j-1)(j+1)} + \frac{n-1}{2n} \right]$$

$$= \frac{2n}{(n-1)(n-2)} \cdot \left[\frac{1}{n} + \sum_{j=2}^{(n-1)/2} \frac{1}{j} \right] \quad \text{(similarly).}$$