

Computer-Intensive Randomization in Systematics

Mark E. Siddall

*Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street,
New York, New York 10024*

Accepted December 12, 2000

There has been a sort of cottage industry in the development of randomization routines in systematics beginning with the bootstrap and the jackknife and, in a sense, culminating with various Monte Carlo routines that have been used to assess the performance of phylogenetic methods in limiting circumstances. These methods can be segregated into three basic areas of interest: measures of support such as bootstrap, jackknife, Permutation Tail Probability, T-PTP, and MoJo; measures of how well independent data are correlated in a phylogenetic framework like PCP for coevolution and Manhattan Stratigraphic Measure (MSM) for stratigraphy; and simulation-based Monte-Carlo methods for ascertaining relative performance of optimality criteria or coding methods. Although one approach to assessing cospeciation questions has been the randomization of, for example, hosts and parasite trees, it is well established that in questions that are of a correlative type, the association themselves are what should be permuted. This has been applied to Brooks' parsimony analysis previously and here to the recent reconciled tree approach to these questions. Although it is debatable whether the extrinsic temporal position of a fossil can stand as refutation of intrinsic morphological character-based cladograms, one can, nonetheless, determine the strength and significance of fit of stratigraphic data to a cladogram. The only method available in this regard that has been shown to not be biased by tree shape is the MSM and modifications of that. Another

similar approach that is new is applied to evaluating the historical informativeness of base composition biases. Incongruence length difference tests too are essentially correlative in nature and comparing the behavior of "perceived" partitions to randomly determined partitions of the same size has become the standard for interpreting the relative conflict between differently acquired data. Unlike the foregoing, which make full use of the observed structure of the data, Monte Carlo methods require the input of parameters or of models and in that sense the results tend to be lacking in verisimilitude. Nonetheless, these kinds of questions seem to have been those most widely promulgated in our field. The well-established theoretical proposition that parsimony has problems with adjacent long-branches was of course illustrated through such methods, much to the concern and angst of systematists. That likelihood later was shown to perform worse than parsimony when those long branches might repel each other has generated less concern and angst. But then many such circumstances can be divined, like the "short-branch-mess" problem wherein likelihood has difficulty placing just a single long branch. Overall, then, in the interpretation of these or any other Monte Carlo issues it will be important to critically examine the structure of the modeled process and the scope of inferences that can be drawn therefrom. Modeling situations that are bound to yield results favorable to only one approach

(such as unrealistic even splitting of ancestral populations at unrealistically predictable times in examination of the coding of polymorphic data) should be viewed with great caution. More to the point, since history is singular and not repeatable, the utility of statistical approaches may itself be dubious except in very special circumstances—most of the requirements for stochasticity and independence can never be met. © 2001 The Willi

Hennig Society

The scientist must be the judge of his own hypotheses, not the statistician.

(Edwards, 1972, p. 34)

BACKGROUND

Types of randomization routines fall essentially into two broad groups. Those that constitute tests of association are called “approximate randomization” methods. A test of correlation, for example, is an approximate randomization method. For questions of correlation the empirical question usually is formulated as: “Is the association between X and Y stronger than we would expect if they were otherwise associated?” So, for example, consider height and weight in three subjects: (4 feet, 70 lbs), (5 feet, 130 lbs), (6 feet, 190 lbs), where $r = 1.000$. “Otherwise associated” can be enumerated in its entirety: (4 feet, 70 lbs), (5 feet, 190 lbs), (6 feet, 130 lbs); or (4 feet, 190 lbs), (5 feet, 130 lbs), (6 feet, 70 lbs); or (4 feet, 130 lbs), (5 feet, 70 lbs), (6 feet, 190 lbs); or (4 feet, 190 lbs), (5 feet, 70 lbs), (6 feet, 130 lbs); or (4 feet, 130 lbs), (5 feet, 190 lbs), (6 feet, 70 lbs). None of these will yield an absolute value of r that is greater than or equal to 1.00; thus the P value is $1/6$ or 0.167. Another kind of test for association is analogous to Fisher’s exact test. Suppose you had two binary variables like door open (O)/door closed (C) and windy (W)/still (S) and three observations,

door O O O C
wind W S W S,

and you aim to assess whether or not there is a nonrandom association between having the door open and it being windy. The expected frequency of finding the door open is 0.75, and that of finding wind is 0.50, such that the expected frequency of finding the door

open and a wind then is 0.375 or 1.5 times out of 4. But is finding 2 of 4 co-occurrences of open door and wind sufficient to conclude that there is a nonrandom association? There are 576 ways in which the observations can be associated. By evaluating all of them one could find that the answer is no. As the number of observations increases in a test of association, be that correlation or otherwise, the number of possible permutations increases geometrically and exact solutions become impractical. Approximate randomization methods circumvent this difficulty by randomly sampling a reasonable number of the possible associations. In practice the number of randomizations is set to some arbitrarily large value like 100 or 1000, though ideally the routine should be run until additional bouts fail to change the P value. The advantages of approximate randomization methods are that they assume no distribution of the data whatsoever and that the data do not even need to satisfy independent and identical distribution requirements for the inference to be valid but then extrapolating inferences beyond the data in hand is not appropriate (Noreen, 1989). Uses of approximate randomization in phylogenetics include the Permutation Tail Probability (PTP) (Faith and Cranston, 1991), the Phylogenetic Covariance Probability (PCP) test for cospeciation (Siddall, 1996b), the incongruence length difference test (Farris *et al.*, 1996), and the Manhattan Stratigraphic Measure (MSM) test of fit of stratigraphic data to a tree (Siddall, 1998a). Whether or not it is reasonable to interpret these as statistical measures is more complex than is implementing the procedures.

The second group of randomization routines are Monte Carlo methods. Monte Carlo methods employ some model for the purposes of generating data. In all such methods the inference made is only as valid as is the model used and how relevant the model is to the question being asked. Usually, these methods are designed to ascertain how well a model performs, such as a model of climate change in relation to increasing carbon dioxide emissions. In Monte Carlo applications the method of inference usually is predetermined, though this is not always the case in phylogenetic applications. Huelsenbeck’s assessment of the effect of long-branch attraction on the performance of various phylogenetic methods (1995; and see Huelsenbeck and Hillis, 1993) used Monte Carlo simulation of data sets according to a model of nucleotide change, as have assessments of long-branch repulsion (Siddall, 1998b).

The use of MoJo (Wenzel and Siddall, 1999) also is a Monte Carlo technique, though in a loosely interpreted way.

The manner in which the results of a randomization method are best interpreted depends on the question being asked, whether the method employed legitimately answers that question, and whether the question is even reasonable. Invariably, the term “confidence” should be avoided by all practitioners of statistical tests when they are employed in a descriptive sense. Conflation of prediction and explanation is common enough, though it is abundantly clear that they are not logically transitive concepts (Scheffler, 1957). That the use of a *t* test or an approximate randomization procedure can well describe the magnitude of the difference in size of fish in two lakes, and the significance (departure from randomness) of that difference, carries no predictive value because there need not be any other fish in the lakes to bother making predictions about. That there are no other fish in the lakes does not deny the validity of being interested in describing differences between those fish that we do have. But this is quite a bit different than the priorities of a toothbrush manufacturer who is confident that her company’s financial resources can tolerate a 5% error rate leading to lawsuits, but then being only 95% confident should be considered inadequate by an elevator manufacturer. The indiscriminate use of confidence in matters that are descriptive, not predictive, usually fails to adequately reflect any sense of *in what* one is confident. This is particularly problematic in systematic applications. It is disconcerting that following Felsenstein’s (1985) description of the bootstrap as a way to get at “confidence limits” few have asked “confidence in what, exactly?” Confident that if one sampled those characters again, one would obtain the same tree? Tautological. Confident, having sequenced 18S rDNA, that cytochrome *c* oxidase I will give the same tree? Far-fetched. Confident that this is the true tree? Hardly. Confident that one’s sample is adequate to answer the question of relationships? Perhaps. But there is no need for positivistic appeals to unknowable “truth.” At best, most of these measures can be interpreted only in terms of “stability” or support. This can only have nonnebulous meaning if one can defend the value of stability (specifically, stability to what) and what is meant by support.

There are two fundamental ways in which randomization has been applied to issues in systematics. Applications of randomization routines in relation to goodness of fit measures are interesting in terms of assessing the behavior of these measures (e.g., Klassen *et al.*, 1991). However, comparing an observed measure of fit to one that is obtained from randomly generated data (Klassen *et al.*, 1991; Meier *et al.*, 1991) is no more meaningful than is assessing a population difference between fish in two lakes with those obtained from randomly generated fork-lengths that would always yield a significant *P* value no matter how the data were structured. Those who have expressed concern at the monotonic decrease in the magnitude of the consistency index (Kluge and Farris, 1969) as more characters are added fail to note that the same monotonic decrease obtains for Pearson’s *r*. If the absolute value of the measure of fit ever was intended to be taken as an indication of the quality of an analysis, then we should be particularly pleased with correlation analyses that have only two data points because in these $|r|$ is guaranteed to be 1.00. But we are not, any more than we are with a consistency index of 1.00 for a three-taxon statement. The use of randomization methods to adjust measures of fit is simply a misunderstanding of what those measures are conveying, and they will not be considered further.

ASSOCIATIONS

Most phylogeneticists are interested in more than the tree itself as an end point with no other meaning than how to name taxa. Often a tree can inform us about character evolution, ecological associations, or other issues of correspondence such as placement in the fossil record. The first problem facing a systematist is how to measure the phenomenon in question in a reasonable and unbiased way. The second is how to assess how meaningful the measure is. Irrespective of the particular question at hand, it should be rather clear that these are questions of correlation, covariation, correspondence, or association and if there is to be some manner of determining how meaningful such a correspondence is, that will be very much like statistical questions of correlation are. By way of introduction to these problems, consider a rather simple question

of allometric association between height and weight (Fig. 1) using Fisher's product-moment statistic r . In one example there are 10 observations and the magnitude of r is 0.794; in the other there are two observations and the magnitude of r is 1.000. At first, then, it would appear that there is a better correspondence in the latter. This is naive of course because it is not the magnitude of the statistic that is meaningful. The magnitude of the statistic simply expresses how well the hypothesis (the line) explains the association of the variables. What is needed, then, is some way to assess how meaningful the magnitude is. One could imagine, for example, "How well does this explanatory hypothesis explain some other set of data," which would be uninteresting to us, as we are concerned only with how well this hypothesis explains these data. It is for this reason (Noreen, 1989), for example, that randomly generating data points according to a Monte Carlo method like parametric bootstrapping (Huelsenbeck *et al.*, 1996c) is meaningless. One could also imagine randomly creating lines and seeing how well they fit the data, but these are guaranteed to be worse than the line we have and so these too are uninteresting, much as the uses of randomly generated trees in character evolution or host-parasite coevolution (Page, 1996) have dubious meaning. The manner in which we assess the significance of fit of associated variables is to randomize the associations because it is the associations that are at stake. For our data points in Fig. 1, this is straightforward. Although the data set composed of

two points (Fig. 1b) has a larger magnitude of r , any and all associations of those variables (63 in., 115 lbs and 68 in., 152 lbs; or 68 in., 115 lbs and 63 in., 152 lbs being the complete set) would resolve an equally large absolute value of r such that $P = 1.00$. The correlation between height and weight for the 10 observations (Fig. 1a), though lower in magnitude, is more significant ($P = 0.008$). That is, in only 0.8% of the possible reassociations of the variables could we achieve a stronger relationship than the one we observe. Tests of association, though certainly statistical, are not ampliative. That a significant fit is concluded does not mean that the addition of more data will not deny this. After all, we are interpolating here, not extrapolating. In addition, finding a significant association is just that, an association; it is not an expression of causality. Finding a strong correspondence between rainfall in the Pyrenees and the price of rhubarb in Nebraska hardly means that the rain in Spain grows rhubarb on the plain.

PTP

One of the first applications of approximate randomization was Archie's (1989) approach to the relative covariations of character information in a cladistic data matrix. Although this has become better known as the Permutation Tail Probability method of Faith and Cranston (1991), they are identical. The question, it

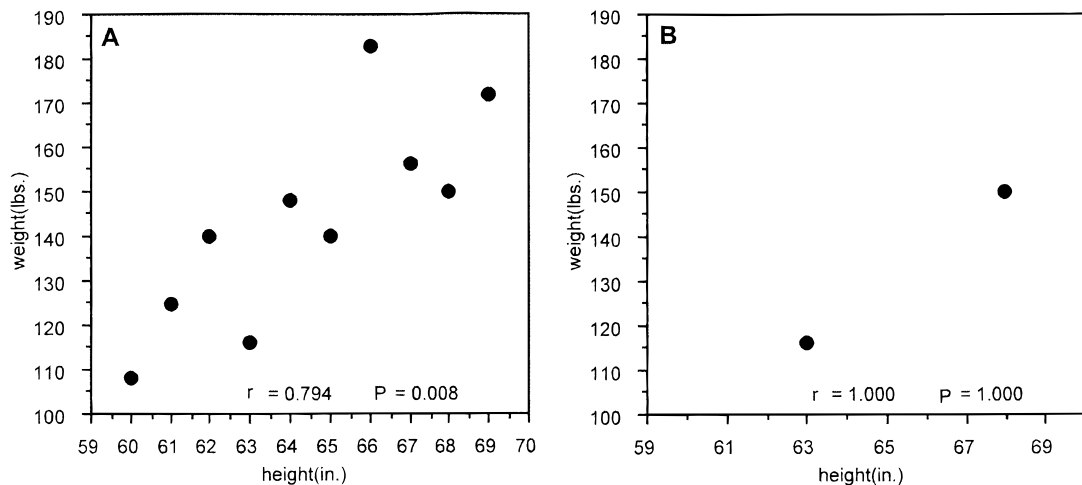


FIG. 1. A typical consideration of correlated variables underscoring the notion that the magnitude of a metric like r is insufficient to express its meaning, unless one also evaluates its departure from what could be expected if the variables were otherwise associated (P).

would seem, is “are character states of different characters randomly or nonrandomly associated in the terminal taxa?” and the method proceeds by way of randomly reassigning the states in each character to the taxa (Fig. 2) and then calculating the most parsimonious tree length from the matrix *schr permuted as follows:

```
for(c = 0; c < chr; c++){
  for(t = 0; t < tax; t++){
    stateholder = schr[chr*t + c];
    schr[chr*t + c] = schr[chr*(k = random(t)) + c];
    schr[chr*k + c] = stateholder;
  }
}
```

In order to assess the utility of this method it is instructive first to examine its behavior in light of truly random matrices. The results of PTP (using the PTP option in Random Cladistics; Siddall, 1996) in relation to a matrix of 6 taxa and 12 characters with four possible states all randomly assigned follow:

```
2 at 29 steps
3 at 30 steps
17 at 31 steps
54 at 32 steps
```

```
24 at 33 steps
"OBSERVED" value = 32
Your PTP-value is: 0.76
```

And, indeed, it would seem that randomly assigned character states appropriately render nonsignificant PTP values. This is expected because it should not matter if one randomly reassigns character states that were randomly assigned in the first place. That PTP behaves appropriately on random data does not mean that it properly assesses departure from randomness when its values are small. Källersjö *et al.* (1992) showed that PTP will give significant values for data sets that lack anything that could be called phylogenetic structure and that in highly structured data it will return nonsignificant values. This caused Faith and Ballard (1994) to reconsider what is meant by “structure” by way of a redefinition that is now “nebulous to the point of inscrutability” (Farris *et al.*, 1994a). The failure of this method of permutation to accomplish what it was intended to measure is now well understood (Källersjö *et al.*, 1992; Farris *et al.*, 1994a; Farris, 1995; Swofford *et al.*, 1996; Carpenter *et al.*, 1998) and its use is not recommended. Except for exceedingly small data sets with inadequate numbers of characters (like those used by Faith and Cranston, 1991) few would have found

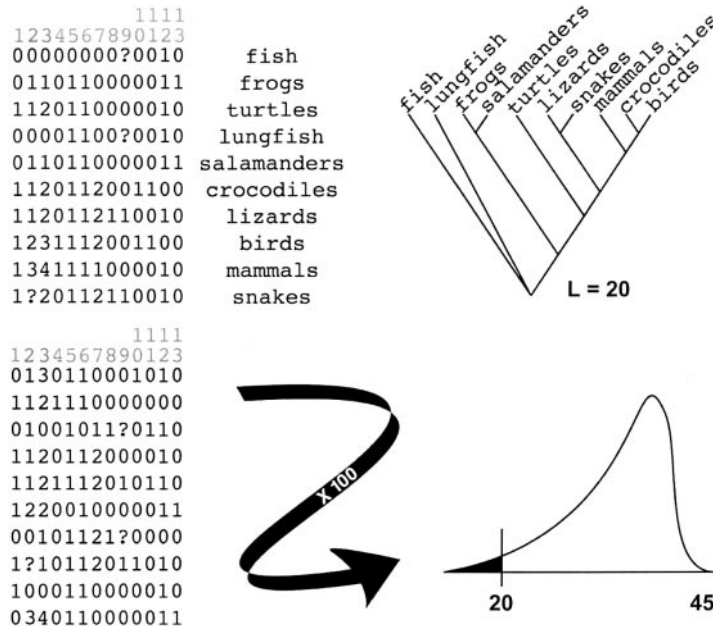


FIG. 2. The Permutation Tail Probability procedure compares the length (L) of the shortest tree on a given set of data (top) with lengths obtained when the states of each character are permuted across taxa (bottom).

insignificant PTP values anyway. Finding that the data fit a most parsimonious tree well is only to be expected because the most parsimonious tree is, by definition, the best fitting hypothesis.

Cospeciation

Comparisons of trees for different sets of taxa can be interesting in phylogenetics if there is some historical association between those taxa as in the case of host–parasite cospeciation or historical biogeography. That is, parasites do not exist without their respective hosts and species must live in areas, areas that are simultaneously occupied by other species. Forms of historical association may well indicate a causally dependent relationship of one set of taxa on some other set of taxa, as in the case of parasite distributions being dependent on their hosts or predators tracking prey through space and time. In other scenarios there is not necessarily an expected causally dependent relationship between taxa, but rather, an expectation that their patterns of divergence are dependent on some third phenomenon such as continental drift (Cracraft, 1988), mountain orogeny (Thorson *et al.*, 1983), changes in river drainage patterns (Mayden, 1988), or two different genes being contained within the same historical lineage (Page, 1993). All of these historical questions suggest that there should be some match between the cladogenetic events observed in each of the associated cladograms. That is, for example, imagine there was a phytophagous insect like a weevil feeding on an ancestral plant species in Southeast Asia before the tectonic collisions that gave rise to what is now Indonesia. If later the plant species dispersed into the newly formed areas, we might expect that the associated insect might also exploit the plant in the newly expanded area (resource tracking). Any length of time or isolation mechanism that could result in speciation of the plant might also be expected to result in speciation of the insect and those cladogenetic events would be correlated. Similarly, a group of parasites, being confined by their respective hosts, would be expected to have cladogenetic events correlated with the same spatio-temporal events of their hosts. If only matters were so simple. What can confound such inferences is that parasites switch hosts, insects can exploit newly encountered plants that are unrelated to those they were

ancestrally associated with, there is nothing to suggest that some species cannot change the number and scope of taxa that they exploit at any given time, taxa can go extinct even though their associate does not, vagile species can disperse, and any given area could have multiple complex overlapping histories that explain contemporary taxonomic compositions. All of these phenomena suggest that one should expect there to be some lack of correlation between any two associated phylogenies.

What measure is suitable as an indication of fit is not well agreed upon, nor is there much consensus as to either the utility of such measures of fit or of how to assess how meaningful is that value of fit. This will not be an extensive review of the competing methods, but, rather, will focus on the application of randomization routines to these specific questions. There are two methods available that employ a randomization protocol for questions of historical association, both of which have unique limitations. Brooks' (1990) parsimony analysis (BPA) requires the experimenter to make a distinction as to which groups being compared are considered to be dependent and which tree is taken to be independent. That is, for example, host phylogeny may be taken to be independent of parasite phylogeny, but the parasites cannot live without their hosts and are, thus, dependent on them. The dependent cladogram of associates can be converted by way of additive binary coding and then optimized on the host cladogram obtaining both a number of steps and a consistency index reflecting the degree of fit of associate phylogeny to the host tree. Each step indicates a historical event that may be correlated cladogenesis, a host switch or a lineage sorting event in which the associate is lost from the host. Because the noncorrelated phenomena of host switching and lineage sorting appear as homoplasy (parallelisms and reversals, respectively), the consistency index (CI) is an indication of the degree of fit (a CI of 1.00 indicates perfect correspondence between parasite and host phylogenies). Siddall (1995a) indicated that the magnitude of such a measure of fit could mislead an investigator because, trivially, any comparison to a host tree with only three taxa is guaranteed to render a perfect fit irrespective of historical phenomena governing the associated distributions. What was needed, then, was a way to ask “could a fit between host and parasite phylogenies

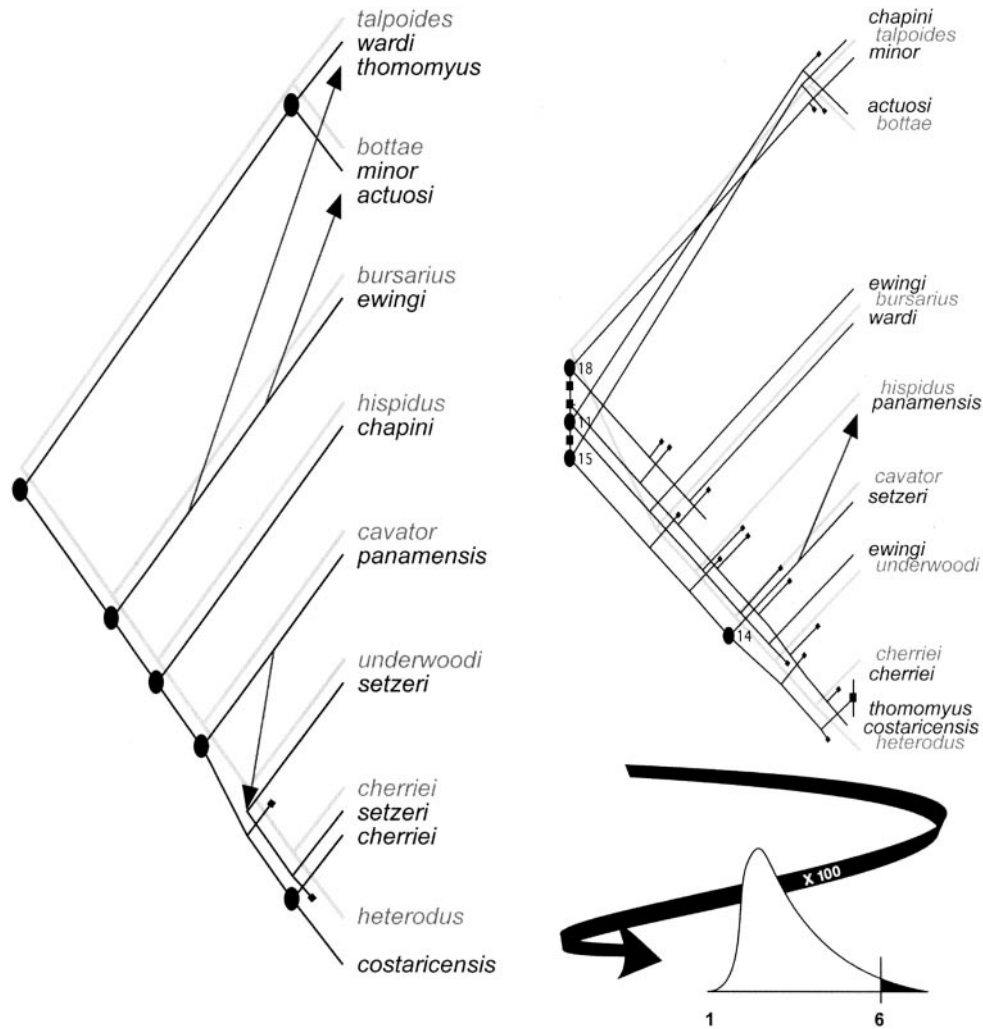


FIG. 3. An approximate randomization procedure to assess the significance of the number of cospeciations found for host and parasite trees (left) should permute associations (top right) and determine the frequency that one could achieve as many or more cospeciation events from those randomized associations (bottom right).

that is as good as this or better have been obtained otherwise?" By "otherwise" it would be inappropriate to compare trees of different shape or composition much as it is inappropriate to assess an r value by way of randomly generated data or randomly considered lines. Because what is at stake is the association of hosts and parasites (or areas and taxa), the appropriate randomization routine is to ask "could I obtain a fit as good or better if the parasites were otherwise associated with these hosts?" This proceeds by randomly reassigning the observed parasites to the available hosts, keeping the number of parasites, number of hosts,

number of associations, and their independently determined phylogenetic histories constant and recomputing the best fit of these reassociated data though BPA (Siddall, 1995a). After doing this many (e.g., 100) times, the frequency that the reassociated data return a fit that is as good or better than the observed associations can be taken as the significance of the original fit (in the original paper I used the word "confidence" and since have been disabused of this rather thoroughly). The greatest difficulty in interpreting this PCP test relates to the well-known flaws in BPA as a reasonable measure of correlated phylogeny. That is, because a

single host switching or lineage sorting event will be counted not only for the switching taxon but also for a suite of its ancestors, BPA will underestimate the goodness of fit between trees. This bias, then, necessarily applies equally to the PCP measure but the effect that this has on underestimating or overestimating the level of significance is not well understood.

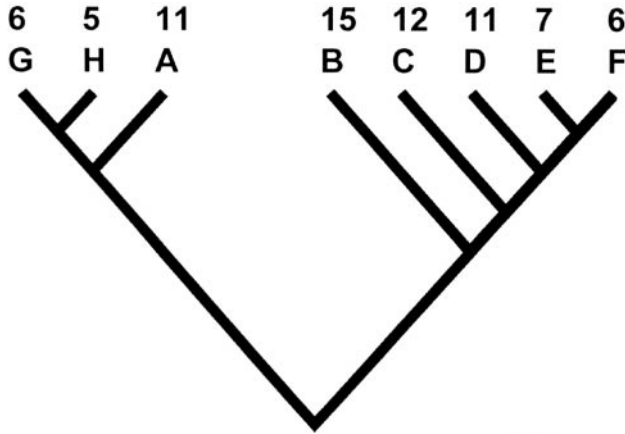
Page's (1994a,b) reconciled tree analysis implemented in TreeMap (Page, 1996) does a better job of evaluating the actual number of correlated cladogenetic events between associated trees. Unlike BPA it is limited to pairwise comparisons of one host tree with one parasite tree as opposed to evaluating the fit of a hosted fauna. Nonetheless, it permits searching for those historical hypotheses of host switching and lineage sorting in relation to the maximum possible number of cospeciations in an unambiguous and appropriately counted way. Page's (1996) approach to determining the significance of an evaluated fit, unfortunately, proceeds by way of assessing the fits of random tree topologies for host and/or parasite taxa while keeping their associations constant. In my opinion, this amounts to asking "if the associated taxa had different histories, could they still be well correlated to each other?" This treats the histories of the species as though they were the alternative possibilities, whereas I would assert that their phylogenetic histories merely are (see also Siddall and Kluge, 1997). Because this is an associative question, it is the associations that should be evaluated. Another way of thinking about this is that it is the disassociation due to host switching and lineage sorting that principally relates to poor fit, and so, if many different kinds of host switching and lineage sorting patterns can return as good a fit or better, then the degree of cospeciation is not significant. The least ad hoc hypothesis of host-parasite relationships found with TreeMap (Page, 1996) for Hafner and Nadler's (1988) data of lice associated with pocket gophers reveals six cospeciation events, three host switching events, and two lineage sorting events (Fig. 3). Randomizing the tree topology for the dependent variable (i.e., parasites) gives a tail probability of 0.01, but the PCP test with BPA, which randomizes associations, gives a tail probability of 0.02. Although Page (1996) does not provide an option for randomized associations, individual files, each with a randomized set of associations, can be created (Fig. 3) and examined with TreeMap (Page, 1996). This indicates that only 2 of 100

such files yield a number of cospeciations as good or better than 6 for this data set (i.e., $P = 0.03$), much as the other measures.

Stratigraphy

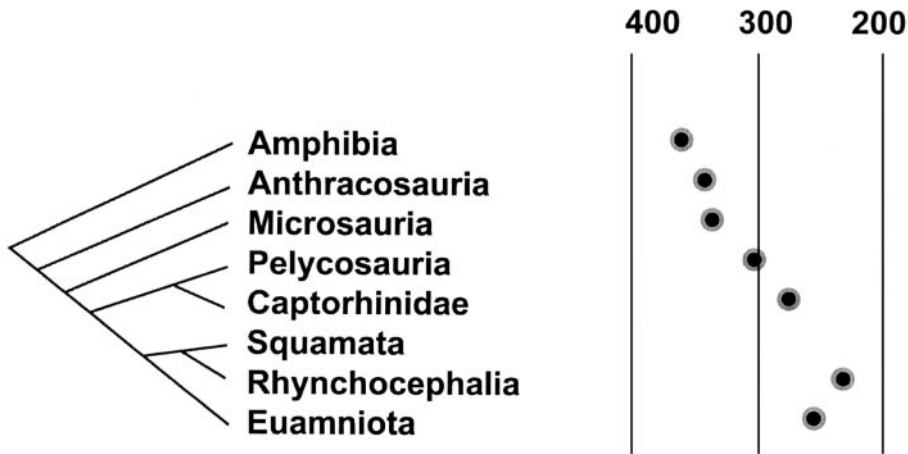
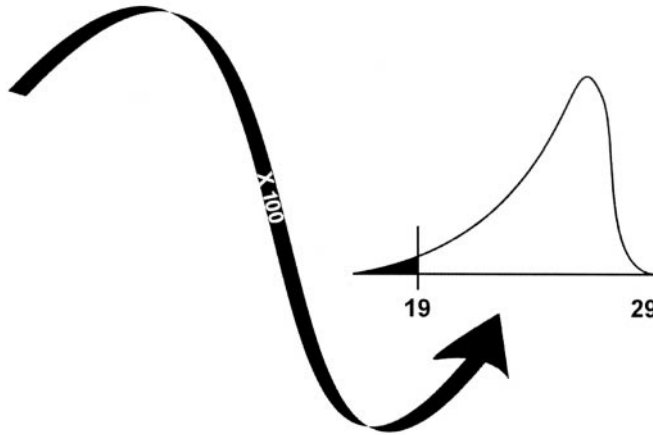
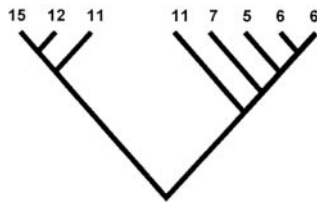
Occasionally an investigator might be interested in how well some other corollary data may be explained by phylogenetic history. Usually these data, whether they be body size, rates of change, base compositions, morphometrics, or age of fossils, fall on a linear continuous scale, which presents a problem when fitting them to a non-Euclidean phylogenetic tree. Gauthier *et al.*, (1988), for example, in fitting stratigraphic ages to phylogenetic trees chose a Spearman correlation statistic of age rank to clade rank but this leaves the value of S open to bias by tree shape such that perfect fits can be obtained only on pectinate trees (Gauthier *et al.*, 1988; Norell and Novacek, 1992). Heulsenbeck's (1994) solution, the Stratigraphic Consistency Index (SCI), though it was claimed otherwise, has a similar bias toward favorable values on pectinate trees except for the trivial situation in which all stratigraphic ages are identical (Siddall, 1995b, 1997a).

It is a little surprising that the use of Manhattan distances as a vehicle for assessing the fit of linear data to trees only recently has been adopted as a means to avoid tree-shape biases (Siddall, 1997a, 1998a; Zelditch *et al.*, 1998) when Manhattan matrices were so central to the origins of cladistic methods. For the case of stratigraphic data [though this has been adopted as well for examining whether base compositions are historically contingent (Sorenson and Siddall, 1997)] a simple protocol addresses both goodness of fit and the significance of that fit (Siddall, 1998a). With the absolute differences in ages for all included T taxa (Fig. 4) establishing the Manhattan stratigraphic distance matrix, and then these values serving as entries in a T by T Sankoff cost matrix, both the length required by the given phylogenetic tree (L_o) and the minimum possible length for these data (L_m) can be determined, the first by simply optimizing the cost matrix on the "known" tree for the taxa (ostensibly based elsewhere on character information) and the second by searching for the shortest tree with Sankoff optimization of the Manhattan distance matrix. The ratio of these two (L_m/L_o) is the Manhattan Stratigraphic Measure, which is bounded between 0 and 1 and has been shown to be



	a	b	c	d	e	f	g	h
a	0	4	1	0	4	5	5	6
b	4	0	3	4	8	9	9	10
c	1	3	0	1	5	6	6	7
d	0	4	1	0	4	5	5	6
e	4	8	5	4	0	1	1	2
f	5	9	6	5	1	0	0	1
g	5	9	6	5	1	0	0	1
h	6	10	7	6	2	1	1	0

$L_m = 10 \text{ My}$ $L_o = 19 \text{ My}$



MSM = 0.83 P = 0.002

FIG. 4. The Manhattan Stratigraphic Measure is the ratio of the minimum possible (L_m) fit of the absolute differences in stratigraphic ages for fossil taxa (top) to the best fit given the association of fossils with strata (L_o). The significance of that fit is determined by reevaluating this ratio for randomized associations of fossils with strata (middle).

properly sensitive to the magnitude of age discrepancies in the fossil record while also not being biased by tree shape (Siddall, 1998a). Again, though MSM takes a value with a particular magnitude, it is clear that any three-taxon statement will always yield $MSM = 1.00$ much as any two-coordinate data set will yield an $|r|$ of 1.00. The significance of the magnitude of MSM, like all such issues of correlation of some variable with a tree, must be addressed by way of permuting the observed ages across the given tree to see whether or not a particular value of MSM is any different than could be obtained by wholly random placement of fossils in strata.

It is not precisely clear how one should interpret poorly fitting variables to a phylogeny. I doubt few would disagree that if there is a good and significant fit, say for fossil data to a phylogenetic hypothesis based on other character information, it is unproblematic to assert that the phylogenetic history of these taxa well explains their distribution in the fossil record. But what of the converse situation? Is a poor fit between the tree and the fossil record so easily interpreted? If there is a poor fit, either the fossil record is in some sense misleading or the phylogenetic tree is. Given that stratigraphic data are notoriously incomplete and subject to regular error from inversions, displacements, and mismeasurement of absolute age, if there is much disagreement between a tree and stratigraphy, surely this would cast doubt on the latter, not the former. However, others have suggested that stratigraphic fit can be used as an optimality criterion in its own right. Huelsenbeck (1994) suggested that the character length of competing trees should be multiplied by the inverse of his SCI. This is not tenable given that SCI has a tree-shape bias but one could imagine a similar approach using the inverse of MSM. Others (e.g., Clyde and Fisher, 1997) suggest that stratigraphic data can be treated as character information in the stratocladistic method, with each time horizon considered being equivalent to one character incorporated into cladistic matrices along side of and equal in value to morphological characters. But for either of these methods to be sensible, stratigraphic ages must be argued to be heritable and their values should be independent (Farris, 1983). I doubt that this can be successfully argued. Thus, for the moment, a poor or nonsignificant MSM value is merely that: a bad fit. No stratigraphic

data have the force of evidence to overturn a hypothesis based on characteristics that are intrinsic properties of the taxa being considered.

Incongruence

Currently, a widely used randomization test assesses the incongruence between two data sets with respect to their independently supporting alternative topologies. Systematists are divided on whether or not different data sets should be combined in phylogenetic analyses (Mickevich and Farris, 1981; Miyamoto, 1985; Kluge, 1989; Bull *et al.*, 1993; Kluge and Wolf, 1993; Chippendale and Wiens, 1994; Huelsenbeck *et al.*, 1994, 1996a, 1996b; De Queiroz *et al.*, 1995; Allard and Carpenter, 1995; Ballard, 1996; Nixon and Carpenter, 1996; Miyamoto and Fitch, 1995). Miyamoto and Fitch (1995, p. 64) argued that “congruent trees obtained from analyses of independent data sets provide the best estimate of the true phylogeny of the group.” Antithetical to this taxonomic congruence approach is that of character congruence in which all data are analyzed together simultaneously (Kluge, 1989; Eernisse and Kluge, 1993). Similarly, Wheeler *et al.* (1993, p. 16) noted that when combining multiple data sets “the foibles of the individual sources of data are not generally shared . . . the only source of shared information is history . . . This is, of course, the rationale and justification for a phylogenetic approach based on total information.” But Huelsenbeck *et al.* (1994; see also Bull *et al.*, 1993, and Huelsenbeck *et al.*, 1996a) advocated combination of data sets only if they could be shown to be homogeneous in advance. These authors did not specify how one was to make such a determination, but Farris *et al.* (1994b, 1995) explained a logical solution.

The number of extra steps on a cladogram, that is, the number of homoplasious transformations, is that amount of change that is not explained by the phylogenetic hypothesis. Or, it is the amount of character information that is incongruent with the tree topology. Given two data sets (A and B), the number of extra steps using A alone (XS_A) or using B alone (XS_B) is the amount of incongruence due to those data sets. Therefore, the number of extra steps found in an analysis that has both of these data sets combined (XS_{comb}) less that from the separate analyses ($XS_A + XS_B$) is the incongruence that is due to the act of combining the

```

Branchellion tor  AcaTTaTatTtTtAtTtTtGGtGcTgAaTcGcaAaTtaTtaGgCaCaTcaAaGatTt
Ozobranchus marg AatTtaTAcTtTtAtaTtTtGGaGcTtGatcTgcAaTaGtagGtataAGtCTyAGaGrt
Myzobdella lugub AcaTTaTatTtTtAtTtTtGGaGcTtGatcTgcAaTaGtagGtataAGtCAaTAAGaAaTt
Calliobdella viv AcaTTaTatTtTtAtTtTtGGaGcTtGatcTgcAaTtaTtaGgCaCaTcaAaGatTt
Piscicola geomet acATTaTaTtTtAtTTTTTgGagcTtGagcAGCAaTaTAGGaaCTtCAaTAAGaATt
Glossiphonia com acATTaTaTtTtAtTTTTTgGagcTtGagcAGCAaTaTAGGaaCTtCAaTAAGaAaTt
Hemicleipsis marg ACaCTaTAcTtTtAtaTtTtGGaGcTtGatcTgcAaTaGtagGaaCaGcTataAGaAaTt
Placobdella para ACaCTaTAcTtTtAtaTtTtGGGcCaTgAaTcGcTataGtTGGtCaGCaATAAGaAaTt
Theromyzon rude ACaTTaTAcTtTtAtaTtTtGGaGcCtGagcGcCaAaTgTaGgCaCaGcCaATAAGaAaTt
Glossiphonia com ACaTTaTAcTtTtAtaTtTtAGGtGcTtGagcTgcCaAaTgTaGgaaCTtGcaATAAGaAaTt
Theromyzon palle ACaTTaTAcTtTtAtaTtTtGGaGcCaTgAaGcGcCaAaTgTaGgaaCaGcCaATAAGaAaTt
Alboglossiphonia ACaTTaTatTtAtaTtTtAGGtGcTtGagcCaGcTataGtagGtCaCaGcCaATAAGaAaTt
Desserobdella ph ACaCTaTAcTtTtAtaTtTtGGGcCtGatcTgcAaTaGtagGaaCaGcCataAGaAaTt
Desserobdella ph ACcCTaTAcTtTtAtaTtTtGGGcCtGatcTgcAaTaGtagGaaCaGcCataAGaAaTt
Desmobdella para ACaTTaTatTtCataTtTtGGNGCaTgAaGcTgCaAaTgTaGgaaCTtGcTataAGaAaTt
Placobdella mont ACaCTaTatTtTtAtaTtTtGGaGcTtGatcGcAaTgTgTaGgaaCaGcCataAGaAaTt
Placobdella papi ACaCTaTAcTtTtAtaTtTtAGGtGcTtGagcTgcCaAaTgTaGgaaCTtGcaATAAGaAaTt
Glossiphonia com ACaTTaTatTtAtaTtTtAGGtGcTtGagcTgcCaAaTgTaGgaaCTtGcaATAAGaAaTt
Glossiphonia whi ACaTTaTatTtAtaTtTtAGGtGcTtGagcTgcCaAaTgTaGgaaCTtGcaATAAGaAaTt
Helobdella fusca ACaTtGtAcTtTtAtaTtTtGGaGcTtGatcGcAaTgTaGgaaCaGcCataAGaAaTt
Haementeria grac ACaTtGtAcTtTtAtaTtTtGGaGcTtGatcGcAaTgTaGgaaCaGcCataAGaAaTt
Haementeria ghil ACaCTaTAcTtTtAtaTtTtAGGtGcCataGgCaGcTataGtagGaaCaGcCataAGaAaTt
Helobdella linea acTtTtAcTtTtAtgTtTtGGgCctGgAaGcTgCtataGtagGaaCTtGcTataAGaAaTt
Helobdella stagn ACaTTaTAcTtTtAtaTtTtGGaGcCaTgAaGcTgCtataGtagGaaCTtGcTataAGaAaTt
Helobdella stagn ACcTTaTAcTtTtAtgTtTtGGgCctGgAaGcTgCtataGtagGaaCTtGcTataAGaAaTt
Helobdella trise acTtTtAcTtTtAtgTtTtGGgCctGgAaGcTgCtataGtagGaaCaGcCataAGaAaTt
Helobdella trans ACtTtAcTtTtAtgTtTtGGgCctGgAaGcTgCtataGtagGaaCaGcCataAGaAaTt
Helobdella elong ACaCTaTAcTtTtAtaTtTtGGtGcCtGgAaGcTgCtataGtagGaaCaGcCataAGaAaTt
Marsupiobdella a ACaCTaTAcTtTtAtaTtTtGGaGcTtGatcGcAaTtaTtaGgCaCaGcCataAGaAaTt
Oligobdella bian ACaTTaTAcTtTtAtaTtTtGGaCaTgAaTcTgcAaTaAaTgGaaCaGcCataAGaAaTt
Helobdella papal ACtTcaTAcTtTtAtgTtTtGGaGcTtGagcTgcCataGtagGaaCaGcCataAGaAaTt
Placobdella para ACaTTaTatTtCataTtTtGGgCcaTgAaTcGcTataGtTGGtCaGCaATAAGaAaTt
Placobdella orna ACaCTtTatTtTtAtaTtTtGGaGcTtGatcGcAaTgTgTaGgaaCaGcCataAGaAaTt
Placobdella pedi ACaCTtTatTtTtAtaTtTtGGaGcTtGatcGcAaTgTaGtagGaaCaGcCataAGaAaTt
Placobdella mole ACaCTtTAcTtTtAtaTtTtGGaGcTtGatcGcAaTgTtGgaaCaGcCataAGaAaTt
Placobdella tran ACaCTaTAcTtTtAtaTtTtGGGcCtGgAaGcTgCtataGtagGaaCaGcCataAGaAaTt
Torix baicalensi ACaTTaTAcTtTtAtgTtTtGGtGcAaTgGcTgCtataGtagGaaCaGcCataAGaAaTt
Theromyzon spp ACaTTaTAcTtTtAtaTtTtGGaGcCaTgAaGcGcCaAaTgTaGgaaCaGcCataAGaAaTt

```

LengthT = 1858
 Length1,2 = 380
 Length3 = 1453

R(LengthT) = 1858
 R(Length1,2) = 1302
 R(Length3) = 538

ILD = 1858-(1453+380) = 25

R(ILD) = 1858-(1302+538) = 18

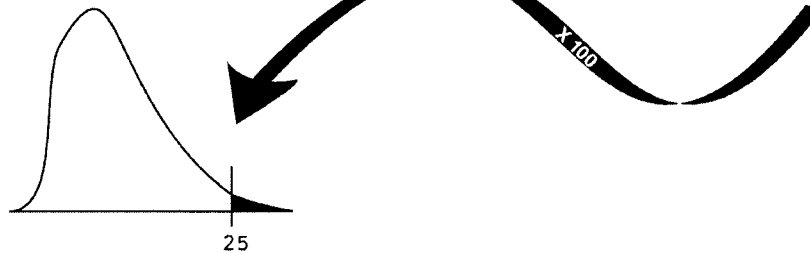


FIG. 5 The significance of the magnitude of the incongruence length difference (ILD) for a particular partition of data (such as positions 1 and 2 versus 3 on left distinguished by case) can be assessed in relation to random partitions of equal size (on right).

data sets together. Whether or not this value of incongruence length difference is significant is readily determined by comparison of its magnitude against random partitions of the total data of equal sizes as the original partition (Fig. 5). However, incongruence and combinability are different things. The proportion of characters in some perceived partition that contribute incongruence might be very small (Siddall, 1997b) and of marginal consequence in light of otherwise overwhelming agreement or mutual corroboration in the two data sets. So, although it may be interesting to know whether or not, for example, two genes exhibit significant disagreement, this is not a justification for excluding or arbitrarily weighting portions of data.

MONTE CARLO MODELS AND METHODS

Certain aspects of the behavior of many methods can be assessed without the use of real data but through controlled randomly generated data instead. The utility of these Monte Carlo approaches necessarily is limited by the specifics of the models used, the parameters that are varied, how realistic modeled conditions might be, and how thoroughly alternative conditions are examined.

Suppose that some method purports to give a reasonable measure of some phenomenon. Two conditions are necessary (though not necessarily sufficient) for that claim to be valid. The method must, of course, take on some meaningful value when the phenomenon in question is present, and it is to this that most simulations are directed. However, the method must also be shown to take on either a low value or an ambivalent value when the phenomenon is known to be absent. In consideration of the utility of PTP, for example, there was modeled (above) a data set composed of randomly assigned character information. Creating a data matrix by randomly choosing from among, say, four states is a form of Monte Carlo simulation. In this, PTP performed as expected because it returned a poor P value (0.76) for the random data. Thus it satisfies this particular condition, though it does not satisfy others. Similar Monte Carlo simulations have proven useful in assessing, for example, the behavior of Huelsenbeck's (1994) SCI (Siddall, 1997a).

A number of parameters can be considered in this

particular Monte Carlo approach such as the number of taxa, the shape of the tree, and the number of possible stratigraphic ages distributed across the tree. When a measure, like SCI, reveals extraordinarily good values for randomly generated (and thus meaningless) data for a particular tree shape there seems to be little reason to avail oneself of the method as an unbiased estimator of fit for real data.

Methods for Coding Polymorphisms

More often, though, Monte Carlo techniques are employed to study the behavior of different methods in certain circumstances. One of the difficulties associated with this approach is that the circumstances that can be simulated in a Monte Carlo design necessarily are stochastic with dubious relevance to understanding real situations of a more deterministic nature. Moreover, that a method might perform well in one set of limiting circumstances does not guarantee that it will do so in some other set of circumstances that are equally cogent. Consider, for example, the question of coding strategies for polymorphic taxa. The use of polymorphic terminal taxa, common to certain data types such as allozyme electromorphs, has been problematic for systematists and there are various methods designed to circumvent analytical difficulties (Mickey and Johnson, 1976; Mickey and Mitter, 1981, 1983; Swofford and Selander, 1981; Buth, 1984; Rogers, 1984, 1986; Swofford and Berlocher, 1987; Nixon and Davis, 1991; Mabee and Humphries, 1993; Mardulyn and Pasteels, 1994; Murphy, 1993; Wiens, 1995). Regarding allozymes, contemporary workers agree that coding the allele as the character is logically and empirically specious, but there is little consensus as to which locus-as-the-character method is to be preferred. Wiens (1995) investigated available methods with the intent of determining the best coding technique and whether or not polymorphic characters contain useful information. Wiens (1995) preferred his frequency method over that of Mabee and Humphries' (1993) scaled Sankoff design but relied on inferential means such as bootstrap values and numbers of characters for this conclusion. A more direct approach would be to actually model the evolution of allozymes, allowing polymorphisms to occur, beginning with a starting population and following through random assortment of alleles

over many generations with mutations and bifurcations of the population in a controlled way.

In the first of these simulations, beginning with one population and ending with four, the rate of cladogenesis was held constant (occurring at the end of every 1000 generations) as was the manner of divergence (ancestral populations being split randomly into two equally sized descendant populations). With the relationships of the four descendant populations being known and predetermined, then, one can investigate how frequently a coding strategy finds the correct tree. Given these circumstances, Wiens' (1995) preference for the frequency approach appears to be well founded (Fig. 6). But this is, after all, a very limiting case.

What happens to performance if divergence time is allowed to vary freely and if the relative proportions in descendant populations also varies freely is telling. Under these conditions, the frequency method performs considerably worse (Fig. 6) and the scaled method of Mabee and Humphries (1993) does considerably better. That is, the "frequency" method did perform better than the alternatives provided that timing of speciation events was constant and that ancestral populations were divided evenly. But then this is not

too surprising given that it is only under these conditions that frequencies are heritable. The principal objection to frequency methods is that frequencies are not necessarily heritable (Farris, 1981; Carpenter *et al.*, 1993; Crother, 1990; Murphy, 1993). Because of the variability of allelic frequencies in space and time, peripheral isolate speciation modes and rapid successions of speciation events will confound their inheritance. Inasmuch as one cannot know what has actually happened in the history of a particular suite of taxa, these simulations offer no direction as to the appropriate choice of method.

The Push and Pull of Long Branches

The problem of interpreting the results of modeling experiments was anticipated by Farris (1986) as it relates to the most commonly employed Monte Carlo design: assessing the performance of various phylogenetic methods under the conditions that might lead to long-branch attraction. Any method "that is consistent under one set of circumstances can be made inconsistent under others; it is only a matter of imagining the circumstances" (Farris, 1986, p. 25; see also Farris, 1983, p. 682; Siddall and Kluge, 1997). Perhaps the best known Monte Carlo simulations are those that have been examined by Huelsenbeck (1995). In this case, changes in nucleotide characters are modeled according to some stochastic process, with the variable of interest being the relative lengths of the five branches on a tree. The results of these simulations (redrawn in Fig. 7) indicate that under these conditions parsimony analysis will find the correct modeled tree less frequently than maximum likelihood provided that two unrelated branches have experienced extraordinarily high rates of change.

This has been taken by many to be rather damning of parsimony analyses in general and as an indication that maximum likelihood is to be preferred. But there are many difficulties associated with such a simple interpretation of these simulations. Edwards (1995), who would otherwise appear to be a staunch advocate of (nonphylogenetic) likelihood methods, indicated that it is only tautological that a method that assumes a stochastic model should perform well in simulations that are stochastic in their design. But, what if the real history of taxa is marked by deterministic phenomena

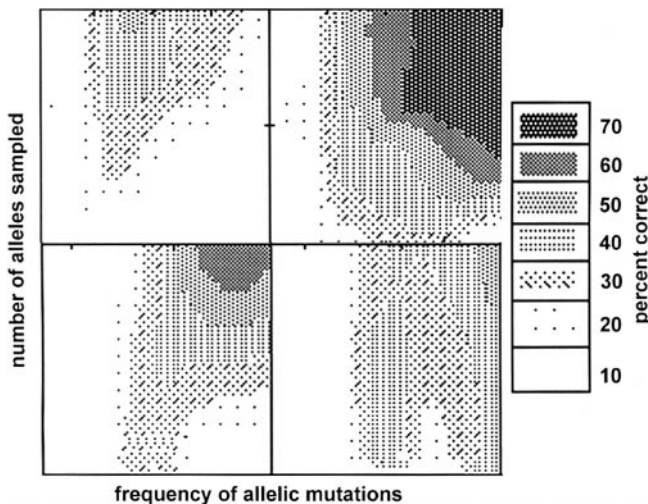


FIG. 6. Performance of two coding strategies in relation to data sets created by simulating the evolution of alleles on a model four-taxon tree while varying the number of loci sampled (5 to 30 on the asymptote) and the frequency of mutation events (increasing on the abscissa). The performance of the two competing methods, the Mabee and Humphries (1993) method on the left versus Wiens' (1995) method on the right, depends on whether speciation events are spaced evenly in space and time (top) or not (bottom).

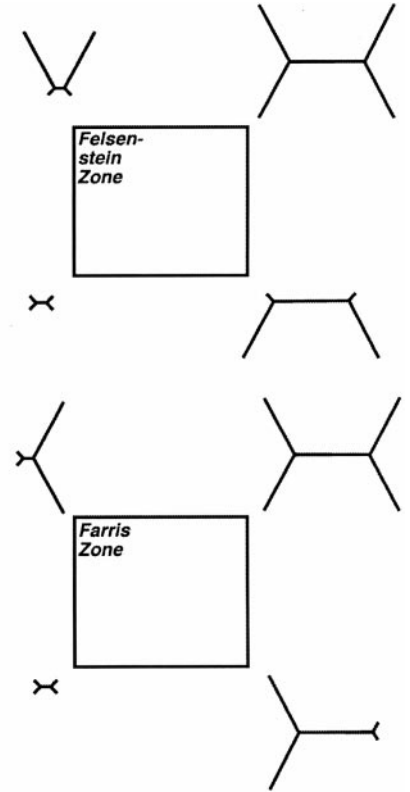
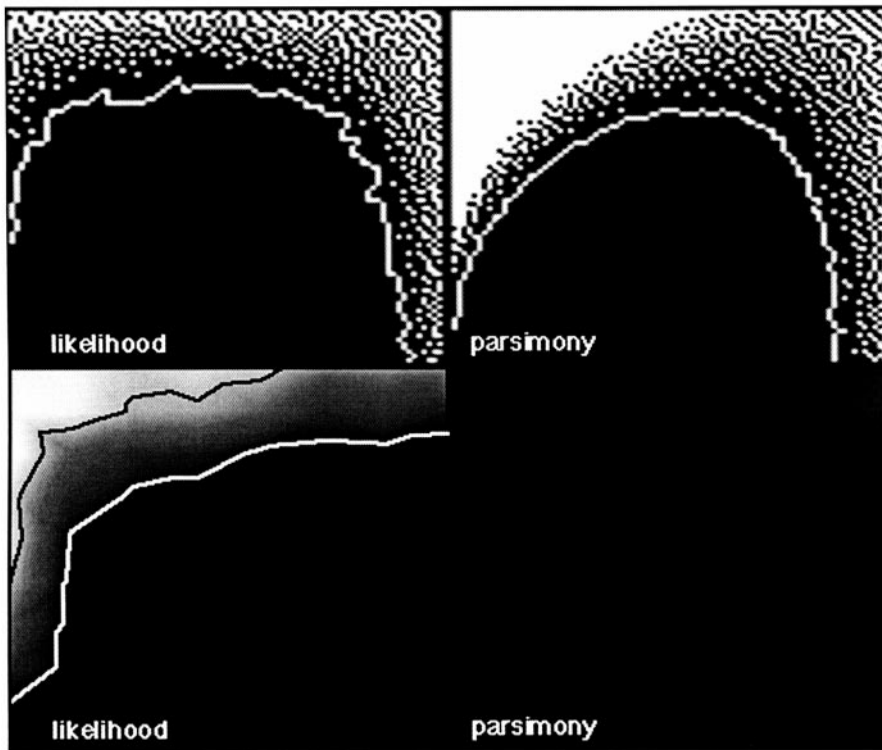


FIG. 7. Interpreting the performance of phylogenetic methods in simulations depends on the specifics of the Monte Carlo technique. When long-branched taxa are separate in model trees (top), parsimony is more subject to long-branch attraction. However, when those long branches are sister taxa in the model tree, likelihood is susceptible to long-branch repulsion.

(Siddall and Kluge, 1997) like, for example, codon usage biases, the vagaries of hydrophobic versus hydrophilic residues, dinucleotide effects or just selection? “If the model employed is not constrained by realism . . . it in fact shows nothing” (Farris, 1986, p. 25).

Other work has underscored how limiting the simulations contrived by Huelsenbeck actually are (Siddall, 1998b; Pol and Siddall, 2001). In similar Monte Carlo fashion, Siddall (1998b) examined a different scope of variable branch lengths with stochastic changes in nucleotides as has been considered previously; in these simulations the two-branch rate was applied to sister taxa in the four-taxon tree. Under these (equally limiting) conditions, maximum likelihood methods performed poorly as the branch lengths of sister taxa were made large (i.e., long-branch repulsion), but parsimony recovered the correct tree in that corner of the parameter space.

More recently (Pol and Siddall, 2001) in the face of criticisms that parsimony only obtains the correct tree in the Farris Zone for the wrong reasons, simulations have focused on more complex tree topologies involving more taxa and also examining the effects of a single long branch (where attraction and repulsion are simply obviated). Clearly as a single long branch becomes longer and longer, to the extent that its nucleotides are randomized relative to the ancestral sequence, any method is going to have increasing difficulty placing this taxon (Fig. 8). However, it is also clear that parsimony is far less susceptible to these problems than is likelihood. Maximum likelihood (using the correct model) has a less than even chance of recovering the modeled tree after an average of 1 substitution per site is reached, whereas parsimony is still recovering the correct tree about 90% of the time (Fig. 8).

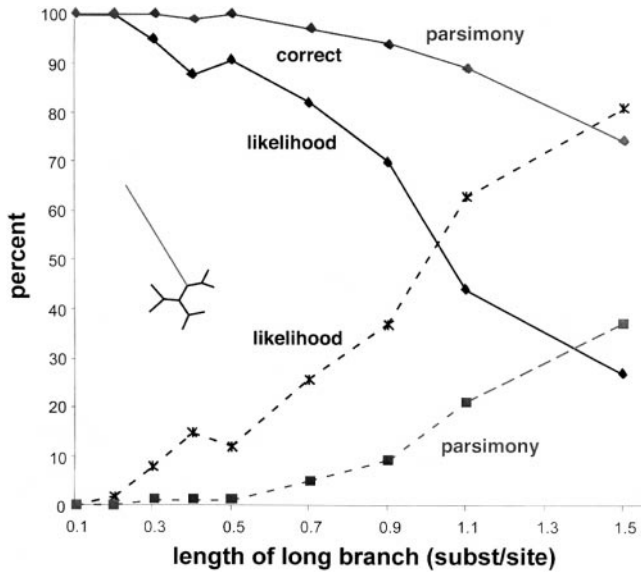


FIG. 8. Simulation of a single long branch on a seven-taxon tree with relative performance of parsimony and likelihood in terms of acquiring the true modeled tree (solid lines) or an incorrect tree (dashed lines) as the long branch is made longer (substitutions per site) (redrawn from Pol and Siddal, 2001).

Base Compositions

The use of weights that correct for base compositional differences “across all taxa” is rare in parsimony-based approaches though such variation in third positions may cause some to eliminate this class of characters. In likelihood applications (e.g., using the F81 or HKY85 models) it is common to incorporate base composition information *from across all taxa* into the estimating function. If base compositions can be shown to be historically contingent, the amelioration of base compositional differences by way of an across all taxa determined weighting function would discard historically relevant information. The determination of some “general” base compositional values across all taxa does not preclude that general value from being inapplicable to the particular subclades of the tree (as admitted by those who favor the LogDet method). To date, no one has developed a simple method for making this determination in a topological sense. In order to examine the historical contingencies of base composition one could develop a method that is similar to those described by Swofford and Berlocher (1987) and undescribed by Weins (1995; and still unpublished by Hillis,

Chippendale and Weins as cited by Weins, 1995) in relation to allelic data. PAUP* (Swofford, 2000) has an import function that will do this automatically provided that the base frequencies are represented in FREQPARS format. In short, the Manhattan distance (d) between each taxon's nucleotide proportions (π) can be determined as follows, for taxa i and j :

$$d_{ij} =$$

$$\frac{|\pi A_i - \pi A_j| + |\pi C_i - \pi C_j| + |\pi G_i - \pi G_j| + |\pi T_i - \pi T_j|}{2}$$

These distances are then entered into a Sankoff matrix (Sankoff and Rousseau, 1975; Sankoff *et al.*, 1976) and optimized on the most parsimonious tree obtaining a length. If base compositions are not historically contingent, then the length so obtained should not differ significantly from that obtained from a nonhistorical distribution of the observed base compositions, that is, from random shuffling of extant base compositions across terminals on the tree. Thus, the tail distribution of optimal tree lengths obtained from repeatedly shuffling base compositions across taxa, in which the length obtained from random associations of this information is less than or equal to the optimized length from observed associations of base compositions, represents the probability that base compositions are not historically contingent. A Macintosh operating system environment Pascal-coded application, HIST π (Sorenson and Siddall, 1997), creates the appropriate randomizations in terms of batch commands executed by PAUP* (Swofford, 2000). Figure 9 shows the results of examining this behavior for the base compositions for the third positions of 13 mitochondrial genes across 14 taxa (Allard *et al.*, 1999), cytochrome *c* oxidase I data for leeches (Siddall and Burrenson, 1998), and sequence characters for the *gag* and *pol* regions of 28 immunodeficiency virus genomes (Mindell *et al.*, 1995). In 3 of these (Fig. 9), none of the 999 random associations of base compositions to terminals resulted in a shorter Sankoff optimization than was obtained using the actual associations. Only 4 of 999 random associations yielded optimizations that were as short or shorter for the mammalian mtDNA data. Thus, in all of these data sets, base compositions are historically contingent and have information that is relevant to the recovery of historical patterns. The downweighting of third positions according to the

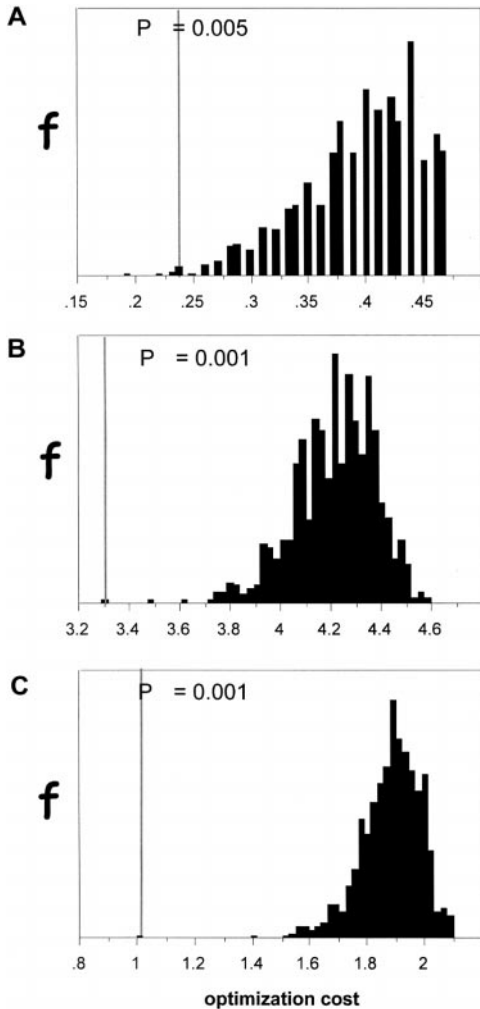


FIG. 9. Assessment of the historical contingency of base compositions across taxa in three data sets. In each of the mammalian mitochondrial genome (A), leech COI (B), and HIV (C) data sets, base composition differences are significantly explained by phylogenetic placement of the taxa relative to a random association of base compositions across taxa. This should preclude the use of all-taxon base composition statistics as in an F81 or an HKY85 model for example.

inverse of base compositions would simply discard a portion of this information content and clearly too the use of across-all-taxa estimates of base compositions in a likelihood function would be untenable.

EPILOGUE

Randomization routines will continue to be developed in the field of phylogenetic systematics. Unfortunately many undoubtedly will continue to view this

as a validation of a statistical approach to phylogeny estimation. The difficulty with statistical methods in matters of historical inference is that history is singular. Without repeated cases (like extracting all the trout in a lake to measure them) or room for abstract generalizations (like a measure of central tendency such as a mean), it is not clear what a statistical measure really has to say. This relegates all randomization methods in systematics to descriptive tools, not inferential tools. No randomization method can offer refutation of a clade; only character evidence can do that, and it must do so directly. In terms of monophyly indices, these randomization routines might be interesting for assessing the stability of a hypothesis to certain kinds of data perturbation, but then it is far from clear how those values are to direct us in our work. What is a good value and what is a bad value will forever be open to subjective determination, and that 65% might be considered “bad” in an analysis of *rbcL* in monocotyledons does not preclude this value from being considered “good” in an analysis of COI of leeches; they are different questions on different data and should never be compared. In any case, nodal values cannot be saved from the nonindependence problem that precludes their statistical interpretation. The use of approximate randomization does not share this problem, but then it is not always clear how departure from randomness should cause one to act. These too simply are tools for describing the relationships between data; a strong and significant fit cannot show causality. And, finally, I suspect that there will be more expositions on how parsimony behaves under certain restrictive and wholly stochastic Monte Carlo modeling regimes. So long as no one seriously thinks that nucleotides are in a stochastic steady state condition across all taxa, we should not be too troubled by such findings and should await the results obtained from modeling the evolution of a femur.

REFERENCES

- Allard, M. W., and Carpenter, J. M. (1996). On weighting and congruence. *Cladistics* **12**, 183–198.
- Allard, M. W., Farris, J. S., and Carpenter, J. M. (1999). Congruence among mammalian mitochondrial genes. *Cladistics* **15**, 75–85.
- Archie, J. W. (1989). A randomization test for phylogenetic information in systematic data. *Syst. Zool.* **38**, 239–252.

- Ballard, J. W. O. (1996). Combining data in phylogenetic analysis. *Trends Ecol. Evol.* **11**, 334–334.
- Brooks, D. R. (1990). Parsimony analysis in historical biogeography and coevolution: Methodological and theoretical update. *Syst. Zool.* **39**, 14–30.
- Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* **42**, 384–397.
- Buth, D. G. (1984). The application of electrophoretic data in systematic studies. *Annu. Rev. Ecol. Syst.* **15**, 501–522.
- Carpenter, J. M., Goloboff, P. A., and Farris, J. S. (1998). PTP is meaningless, T-PTP is contradictory: A reply to Trueman. *Cladistics* **14**, 105–116.
- Carpenter, J. M., Strassmann, J. E., Turillazzi, S., Hughes, C. R., Solis, C. R., and Cervo, R. (1993). Phylogenetic relationships among paper wasp social parasites and their hosts (Hymenoptera: Vespidae; Polistinae). *Cladistics* **9**, 129–146.
- Chippindale, P. T., and Wiens, J. J. (1994). Weighting, partitioning, and combining characters in phylogenetic analysis. *Syst. Biol.* **43**, 278–287.
- Clyde, W. C., and Fisher, D. C. (1997). Comparing the fit of stratigraphic and morphologic data in phylogenetic analysis. *Paleobiology* **23**, 1–19.
- Cracraft, J. (1988). Deep-history biogeography: Retrieving the historical pattern of evolving continental biotas. *Syst. Zool.* **37**, 221–236.
- Crother, B. I. (1990). Is some better than none or do allele frequencies contain phylogenetically useful information? *Cladistics* **6**, 277–282.
- De Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Syst.* **26**, 657–681.
- Edwards, A. W. F. (1972). "Likelihood." Cambridge Univ. Press, Cambridge, UK.
- Edwards, A. W. F. (1995). Assessing molecular phylogenies. *Science* **267**, 253.
- Eernisse, D. J., and Kluge, A. G. (1993). Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Mol. Biol. Evol.* **10**, 1170–1195.
- Faith, D. P., and Ballard, J. W. O., (1994). Length differences topology-dependent tests—A response. *Cladistics* **10**, 57–64.
- Faith, D. P., and Cranston, P. S. (1991). Could a cladogram this short have arisen by change alone?: On permutation tests for cladistic structure. *Cladistics* **7**, 1–28.
- Farris, J. S. (1981). Distance data in phylogenetic analysis. In "Advances in Cladistics I" (V. A. Funk and D. R. Brooks, Eds.), pp. 3–23. N.Y. Bot. Garden, New York.
- Farris, J. S. (1983). The logical basis of phylogenetic analysis. In "Advances in Cladistics" (N. I. Platnick and V. A. Funk, Eds.), Vol. 2, pp. 7–36. Columbia Univ. Press, New York.
- Farris, J. S. (1986). On the boundaries of phylogenetic systematics. *Cladistics* **2**, 14–27.
- Farris, J. S. (1995). Conjectures and refutations. *Cladistics* **11**, 105–118.
- Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1994a). Permutations. *Cladistics* **10**, 65–76.
- Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1994b). Testing significance of incongruence. *Cladistics* **10**, 315–319.
- Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1995). Constructing a significance test for incongruence. *Syst. Biol.* **44**, 570–572.
- Farris, J. S., Albert, V. A., Källersjö, M., Lipscomb, D., and Kluge, A. G. (1996). Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **12**, 99–124.
- Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791.
- Gauthier, J., Kluge, A. G., and Rowe, T. (1988). Amniote phylogeny and the importance of fossils. *Cladistics* **4**, 105–209.
- Hafner, M. S., and Nadler, S. A. (1988). Phylogenetic trees support the coevolution of parasites and their hosts. *Nature* **332**, 258–259.
- Huelsenbeck, J. P. (1994). Comparing the stratigraphic record to estimates of phylogeny. *Paleobiology* **20**, 470–483.
- Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48.
- Huelsenbeck, J. P., and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **42**, 247–264.
- Huelsenbeck, J. P., Bull, J. J., and Cunningham, C. W. (1996a). Combining data in phylogenetic analysis. *Trends Ecol. Evol.* **11**, 152–158.
- Huelsenbeck, J. P., Bull, J. J., and Cunningham, C. W. (1996b). Combining data in phylogenetic analysis—Reply. *Trends Ecol. Evol.* **11**, 335–335.
- Huelsenbeck, J. P., Hillis, D. M., and Jones, R. (1996c). Parametric bootstrapping in molecular phylogenetics: Applications and performance. In "Molecular Zoology: Advances, Strategies, and Protocols" (J. D. Ferraris and S. R. Palumbi, Eds.), pp. 19–46. Wiley-Liss, New York.
- Huelsenbeck, J. P., Swofford, D. L., Cunningham, C. W., Bull, J. J., and Waddell, P. J. (1994). Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis. *Syst. Biol.* **43**, 288–329.
- Källersjö, M., Farris, J. S., Kluge, A. G., and Bult, C. (1992). Skewness and permutation. *Cladistics* **8**, 275–287.
- Klassen, G. J., Mooi, R. D., and Locke, A. (1991). Consistency indexes and random data. *Syst. Zool.* **40**, 446–457.
- Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* **38**, 7–25.
- Kluge, A. G., and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. *Syst. Zool.* **18**, 1–32.
- Kluge, A. G., and Wolf, A. J. (1993). Cladistics: What's in a word? *Cladistics* **9**, 183–199.
- Mabee, P. M., and Humphries, J. (1993). Coding polymorphic data: Examples from allozymes and ontogeny. *Syst. Biol.* **42**, 166–181.
- Mardulyn, P., and Pasteels, J. M. (1994). Coding allozyme data using step matrices: Defining new original states for the ancestral taxa. *Syst. Biol.* **43**, 567–572.
- Mayden, R. L. (1988). Vicariance biogeography, parsimony, and evolution in North American freshwater fishes. *Syst. Zool.* **37**, 329–355.
- Meier, R., Kores, P., and Darwin, S. (1991). Homoplasy slope ratio—A

- better measurement of observed homoplasy in cladistic analyses. *Syst. Zool.* **40**, 74–88.
- Mickevich, M. F., and Farris, J. S. (1981). The implications of congruence in *Menidia*. *Syst. Zool.* **30**, 351–369.
- Mickevich, M. F., and Johnson, M. S. (1976). Congruence between morphological and allozyme data in evolutionary inference and character evolution. *Syst. Zool.* **25**, 260–270.
- Mickevich, M. F., and Mitter, C. M. (1981). Treating polymorphic characters in systematics: A phylogenetic treatment of electrophoretic data. In “Advances in Cladistics I” (V. A. Funk and D. R. Brooks, Eds.), pp. 45–58. N.Y. Bot. Garden, New York.
- Mickevich, M. F., and Mitter, C. M. (1983). Evolutionary patterns in allozyme data: A systematic approach. In “Advances in Cladistics II” (N. I. Platnick and V. A. Funk, Eds.), pp. 169–176. Columbia Univ. Press, New York.
- Mindell, D. P., Shultz, J. W., and Ewald, P. W. (1995). The AIDS pandemic is new, but is HIV new? *Syst. Biol.* **44**, 77–92.
- Miyamoto, M. M. (1985). Consensus cladograms and general classifications. *Cladistics* **1**, 186–189.
- Miyamoto, M. M., and Fitch, W. M. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* **44**, 64–76.
- Murphy, R. W. (1993). The phylogenetic analysis of allozyme data: Invalidity of coding alleles by presence/absence and recommended procedures. *Biochem. Syst. Ecol.* **21**, 25–38.
- Nixon, K. C., and Carpenter, J. M. (1996). On simultaneous analysis. *Cladistics* **12**, 221–241.
- Nixon, K. C., and Davis, J. I. (1991). Polymorphic taxa, missing values and cladistic analysis. *Cladistics* **7**, 233–241.
- Noreen, E. W. (1989). “Computer-Intensive Methods for Testing Hypotheses: An Introduction.” Wiley, New York.
- Norell, M., and Novacek, M. (1992). The fossil record and evolution: Comparing cladistic and paleontologic evidence for vertebrate history. *Science* **255**, 1690–1693.
- Page, R. D. M. (1993). Genes, organisms, and areas—the problem of multiple lineages. *Syst. Biol.* **42**, 77–84.
- Page, R. D. M. (1994a). Parallel phylogenies—Reconstructing the history of host–parasite assemblages. *Cladistics* **10**, 155–173.
- Page, R. D. M. (1994b). Maps between trees and cladistic-analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* **43**, 58–77.
- Page, R. D. M. (1996). TreeMap. Software available from the author.
- Pol, D., and Siddall, M. E. (2001). Biases in maximum likelihood and parsimony: A simulation approach to a ten-taxon case. *Cladistics*, in press.
- Rogers, J. S. (1984). Deriving phylogenetic trees from allele frequencies. *Syst. Zool.* **33**, 52–63.
- Rogers, J. S. (1986). Deriving phylogenetic trees from allele frequencies: A comparison of nine genetic distances. *Syst. Zool.* **35**, 297–310.
- Sankoff, D., and Rousseau, P. (1975). Locating the vertices of a Steiner tree in an arbitrary metric space. *Math. Prog.* **9**, 240–246.
- Sankoff, D., Cedergren, R. J., and Lapalme, G. (1976). Frequency of insertion–deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *J. Mol. Evol.* **7**, 133–149.
- Scheffler, I. (1957). Explanation, prediction and abstraction. In “Philosophy of Science” (A. Danto and S. Morgenstern, Eds.), pp. 274–287. The World Publishing Co., New York.
- Siddall, M. E. (1995a). Phylogenetic covariance probability: Confidence and historical associations. *Syst. Biol.* **45**, 48–66.
- Siddall, M. E. (1995b). Stratigraphic consistency and the shape of things. *Syst. Biol.* **45**, 111–115.
- Siddall, M. E. (1996). Random Cladistics. Department of Biology, University of Toronto, Toronto, Ontario, Canada.
- Siddall, M. E. (1997a). Stratigraphic indices in the balance: A reply to Hitchin and Benton. *Syst. Biol.* **46**, 569–573.
- Siddall, M. E. (1997b). Prior agreement: Arbitration or arbitrary. *Syst. Biol.* **46**, 766–770.
- Siddall, M. E. (1998a). Stratigraphic fit to phylogenies: A proposed solution. *Cladistics* **14**, 201–208.
- Siddall, M. E. (1998b). Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris zone. *Cladistics* **14**, 209–220.
- Siddall, M. E., and Bureson, E. M. (1998). Phylogeny of leeches (Hirudinea) based on mitochondrial cytochrome c oxidase subunit I. *Mol. Phylogenet. Evol.* **9**, 156–162.
- Siddall, M. E., and Kluge, A. G. (1997). Probabilism and phylogenetic inference. *Cladistics* **13**, 313–336.
- Sorenson, M. D., and Siddall, M. E. (1997). Hist π software for determining the association of base compositions to tree structure. Museum of Zoology, University of Michigan, Ann Arbor, MI.
- Swofford, D. L. (2000). PAUP* 4.0: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer, Sunderland, MA.
- Swofford, D. L., and Berlocher, S. H. (1987). Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. *Syst. Zool.* **36**, 293–325.
- Swofford, D. L., and Selander, R. B. (1981). BIOSYS-1: A FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics. *J. Hered.* **72**, 281–283.
- Swofford, D. L., Thorne, J. L., Felsenstein, J., and Wiegmann, B. M. (1996). The topology-dependent permutation test for monophyly does not test for monophyly. *Syst. Biol.* **45**, 575–579.
- Thorston, T. B., Brooks, D. R., and Mayes, M. A. (1983). The evolution of freshwater adaptation in stingrays. *Nat. Geog. Soc. Res. Rep.* **15**, 663–694.
- Wenzel, J. J., and Siddall, M. E. (1999). Noise. *Cladistics* **15**, 51–64.
- Wheeler, W. C., Cartwright, P., and Hayashi, C. Y. (1993). Arthropod phylogeny: A combined approach. *Cladistics* **9**, 1–39.
- Wiens, J. J. (1995). Polymorphic characters in phylogenetic systematics. *Syst. Biol.* **44**, 482–500.
- Zelditch, M. L., Fink, W. L., Swiderski, D. L., and Lundrigan, B. L. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny: A reply to Rohlf. *Syst. Biol.* **47**, 159–167.