Relative Apparent Synapomorphy Analysis (RASA) I: The Statistical Measurement of Phylogenetic Signal

James Lyons-Weiler, Guy A. Hoelzer,* and Robin J. Tausch*†*

*University of Nevada, and †USDA Forest Service, Reno, Nevada

We have developed a new approach to the measurement of phylogenetic signal in character state matrices called relative apparent synapomorphy analysis (RASA). RASA provides a deterministic, statistical measure of natural cladistic hierarchy (phylogenetic signal) in character state matrices. The method works by determining whether a measure of the rate of increase of cladistic similarity among pairs of taxa as a function of phenetic similarity is greater than a null equiprobable rate of increase. Our investigation of the utility and limitations of RASA using simulated and bacteriophage T7 data sets indicates that the method has numerous advantages over existing measures of signal. A first advantage is computational efficiency. A second advantage is that RASA employs known methods of statistical inference, providing measurable sensitivity and power. The performance of RASA is examined under various conditions of branching evolution as the number of characters, character states per character, and mutations per branch length are varied. RASA appears to provide an unbiased and reliable measure of phylogenetic signal, and the general approach promises to be useful in the development of new techniques that should increase the rigor and reliability of phylogenetic estimates.

Introduction

The determination that a particular data set contains phylogenetic signal should be an important first step in phylogenetic analyses, because a set of optimal (albeit spurious) trees can be inferred under any criterion from random (i.e., uninformative) data. While the arsenal of quantitative methods available to the practicing phylogenetic systematist holds an array of descriptive statistics (e.g., as described in Forey et al. 1992), and resampling methods (e.g., bootstrap; Felsenstein 1985) aimed at addressing the issue of certainty in phylogenetics, the sensitivity and power of these methods are largely unknown. Power here is used in the statistical sense (1 − power [Type II error]), and should not be conflated with the usage by Penny et al. (1993) and Hillis, Hulsenbeck, and Swofford (1994), who write of algorithmic power and efficiency. The probability of failing to reject a false null hypothesis is difficult to measure in the absence of stated nulls, independent probability distributions, error terms, and test statistics and their critical values. As the true history of past diversification for a given group is unknown, no objective criterion of truth exists for phylogenetic trees; thus, confidence is a difficult thing to assess in phylogenetics. Some recent methods include sites tests (Kishino-Hasagawa 1989; Steel, Lockhart, and Penny 1995), a GC-modified test (Steel, Lockhart, and Penny 1993), the complete and partial bootstrap (Zharkikh and Li 1995), and a goodness-of-fit test of pattern frequencies (Czelusniak and Goodman 1995). These approaches are largely aimed at choosing among alternative hypotheses of phylogeny as may be inferred from available data, or testing for specific sources of misleading pattern. Exactly how to report confidence in trees remains a nascent field of study (Archie 1989; Goloboff 1991; Huelsenbeck 1991; Mooers 1995). However, it is analytic that reliable trees should be found with a higher frequency when data used to infer the relationships contain phylogenetic signal (= natural cladistic hierarchy = character covariation; Archie 1989; Faith and Cranston 1991). The ideal measure of phylogenetic signal should be able to discern between pattern due to convergence (Faith 1989), and nonindependence of character state evolution. Methods developed thus far to measure phylogenetic signal include tree-length distribution (TLD) central moments (e.g., $g1$; Hillis 1991; Huelsenbeck 1991), and permutation tail probability tests (PTP tests; Archie 1989; Faith and Cranston 1991). Steel, Lockhart, and Penny (1993) found that the PTP and bootstrap could provide misleading results with random data. Furthermore, complete enumeration of these tests requires knowledge of the length of the most parsimonious tree (mpt) and of the TLD for the central moments statistics ($g1$; Hillis 1991). This fact makes exact realization of these measures computationally intensive, as their solutions are only available through tree-building algorithms that involve exhaustive searches for the mpt.

The problem of finding an efficient algorithm to guarantee knowledge of the mpt is known to be an NP-complete problem (Garey and Johnson 1979; Warnow 1993). NP-complete problems are those for which an efficient algorithm is thought to be impossible. The cost of efficient algorithms can, by definition, be described using polynomial terms (Garey and Johnson 1979). An algorithm described by a cost equation that includes any exponent is not considered efficient.

Usual "solutions" to NP-complete problems include heuristic approximations, randomization steps, and resampling methods (i.e., pseudopolynomial time solutions; Garey and Johnson 1979). However, such "solutions" can carry costs in the form of newly emer-

Abbreviations: RASA, relative apparent synapomorphy analysis; E, informative similarity; t_{RASA}, RASA test statistic; L, number of characters; C, number of character states per character; B, branch length.

Key words: phylogenetic certainty, phylogenetic signal, statistical phylogenetics, NP-complete, polynomial time solutions, bootstrap, PTP.

Address for correspondence and reprints: J. Lyons-Weiler, Graduate Program in Ecology, Evolution and Conservation Biology, University of Nevada, Reno, Nevada 89512-0013. E-mail: weiler@grass.ers.unr.edu.

Mol. Biol. Evol. 13(6):749–757. 1996
© 1996 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

gent properties, such as entry-order sensitivity. Entry-order sensitivity occurs when different solutions can be achieved with the same data set by changing the sequence of input. This property exists for some algorithms (especially those that are nondeterministic sensu Warnow 1993). Some examples include algorithms that perform heuristic searches for the mpt (Maddison 1991), as well as many multivariate ordination and classification analyses (Tausch et al. 1996). With respect to the measurement of phylogenetic signal, entry-order sensitivity has not been shown, but seems to be quite likely for approximate, nondeterministic solutions for *g1* and PTP, especially in light of the fact that biased estimates may exist due to unequal topology probabilities (Simberloff 1987) and deviations of observed null frequencies from Markovian expectation (Losos and Alder 1995). In this paper, we present a method (relative apparent synapomorphy analysis [RASA]) that provides a deterministic measure of phylogenetic signal in polynomial time.

The Philosophical Basis for RASA

This approach to the problem of measuring phylogenetic signal is based on the relationship between the basic unit of currency in phylogenetic systematics (parsimony methods) and the building blocks of a given phylogeny. The universal axiom in maximum-parsimony analyses is that monophyletic groups are deduced from shared character states, or synapomorphies (Hennig 1966; Wiley 1981). Synapomorphies are the currency of the phylogenetic systematist, and are usually inferred in a post hoc fashion (but see Steel, Lockhart, and Penny 1993) after the application of some criterion for tree selection (usually minimum distance). The building blocks of any phylogenetic tree are three-taxon statements (TTS; e.g., Nelson and Platnick 1991), and any phylogeny can be reduced to its component TTSs (Wiley 1981). TTSs have a characteristic that is unique among n-taxon statements: they are the only nontrivial n-taxon statements that have fewer possible items of error than members. In fact, they can represent one and only one possible independent error. The relationship between the synapomorphy and the TTS is that a resolved TTS is an inference made from at least one synapomorphy.

The basic tenet of the Hennigian auxiliary principle (Hennig 1966) is that the lowest possible number of changes along all anagenetic lineages best represents the pattern of evolutionary change for all characters for all taxa everywhere. While this principle may hold true most of the time, its universal truth can be rejected on the basis of parsimony (it is simpler to envisage a world where evolution occurs at least sometimes in a nonparsimonious fashion) and the case with which the principle is falsified for some characters (e.g., transitional changes at some positions in DNA sequences).

A reasonable relaxation of Hennig's auxiliary principle permits the position that any two identical character states in different taxa may have arisen independently, and therefore may not be true synapomorphies (i.e., "assume convergence [noise] until sufficient evidence is found to reject the null"). A TTS based on a particular shared character state may therefore also represent an error. Therefore, a particular character state shared between two taxa to the exclusion of other taxa might most accurately be described as an apparent synapomorphy (i.e., it might represent a convergence). This view is compatible with the fusion between phenetic (similarity) and cladistic approaches made possible by Hadamard conjugations used in spectral analysis (Waddell et al. 1993), through which similarity and character pattern are invertible. The distinction between cladistics and phenetics remains, however, because cladistic error exists in the form of increased (or decreased) similarity that occurs independent of phylogeny.

This approach represents a significant departure from the traditional viewpoint of phylogenetic systematics. Here, it is assumed a priori that each and every potential synapomorphy and any inferred TTS may represent individual errors.

The RASA Algorithm

To measure whether an apparent synapomorphy **S** denotes evolutionary kinship for taxa i and j (or whether it represents an error), the total number of times taxa i and j share **S** to the exclusion of another taxon provides information on how unique that single observed similarity is. Each of m taxon pairs can be assigned a Relative Apparent Synapomorphy score (RAS_{ij}), determined to be

$$\mathrm{RAS}_{i,j} = \sum_{k \in \mathbf{A}}^{N} \sum_{c=1}^{n} r \qquad (1)$$

where **A** is a set of taxa for which character state $c(\mathbf{A}) \neq c(n)_i$ or $c(n)_j$, and N = the Nth taxon such that $i \neq j \neq k$. For the nth character, taxon $k \in \mathbf{A}$, taxa i and $j \in \mathbf{A}$, and $r = 1$ if character state $c(n)_i = c(n)_j \neq c(n)_k$. RAS_{ij} is the number of times a taxon other than i and j has a different character state when i and j share the same character state. Multistate characters need not be binary-recoded.

RAS_{ij} is the sum of the number of times taxa i and j share any character state to the exclusion of another taxon, all characters considered. The *RAS* score conveys information on how much unique similarity exists between two taxa with no redundant information added. In general, for a character to be potentially informative, at least two taxa must share one character state to the exclusion of at least one other taxon. Uninformative characters which have insufficient or excessive variability are irrelevant to the determination of RAS_{ij}. A closely related pair of taxa in a given matrix will share a high *RAS*, as they ought to share more derived states to the exclusion of more taxa than do pairs of taxa that are more distantly related, which will exhibit low *RAS* scores.

RAS_{ij} will tend to increase as a function of the overall similarity of i and j. Let E_{ij} = the number of characters that can support the three-taxon statement $(((i, j), k))$. This phenetic (distance) measure of similarity uses only shared character states that are potentially infor-

mative (= informative similarity) for i and j. E_{ij} is the number of characters involved in the computation of RAS_{ij}.

A plot of RAS_{ij} against E_{ij} usually produces a positive slope (observed sloped; β_{obs}). This slope is the rate at which a cladistic measurement (*RAS*) increases as a function of a phenetic measurement (E). For every unitary increase E, there is (on average) an increase in *RAS* as indicated by β_{obs}. The slope of *RAS* on E is a measure of character covariation in a given data set. Alone, however, the observed slope is uninformative, because a random distribution of character states could also lead to a positive slope through variation in phenetic similarity among taxa, and concomitant variation between *RAS* scores.

Data sets that are uninformative with respect to the phylogenetic relationships between their component taxa (e.g., a random distribution of character states) are expected to have an equiprobable distribution of E among all taxa (E_{null}), and a corresponding equiprobable distribution of RAS (RAS_{null}). To determine the appropriate null slope for comparison to the observed slope, *RAS* must be redistributed as a function of E (eq. 2), and E must be redistributed as a function of *RAS* (eq. 3):

$$\mathrm{RAS}_{\mathrm{null}(i,j)} = \mathrm{RAS}_{i,j} * E_{i,j} \Bigg/ \frac{\sum_{k=1}^{m} E_k}{m}, \qquad (2)$$

$$E_{\mathrm{null}(i,j)} = \mathrm{RAS}_{i,j} * E_{i,j} \Bigg/ \frac{\sum_{k=1}^{m} \mathrm{RAS}_k}{m} \qquad (3)$$

This process is called reciprocal equiprobable redistribution (RER). If there is nonrandom character covariation in the data, RER will deterministically redistribute the structure in the data among all pairs of taxa; if not, noise will be redistributed. Least-squares simple regression of RAS_{null} on E_{null} always results in a line without error, an intercept of zero, and a slope (β_{null}) that serves as the appropriate null to which β_{obs} can be compared with the test statistic (Myers 1990) for homogeneity of slopes (t_{RASA}). The null hypothesis is

H_0: *"There will be no significant difference between β_{obs} and β_{null}."*

The inferential null hypothesis is

"Character states do not covary more than expected from a random distribution of phylogenetic signal across pairs of taxa."

The test statistic requires $(N \times (N - 1)/2) - N - 3$ degrees of freedom (see Appendix), where N is the number of taxa in the matrix. If the slopes do not differ, the null hypothesis cannot be rejected, and the result of a phylogenetic analysis with the data in question may be misleading (e.g., tree-building algorithms will construct a spurious tree from random data). Other approaches which are thought to increase signal-to-noise ratios are not incompatible with the search for signal (see Discussion).

Note that the degrees of freedom are a function of the number of taxon pairs in a given matrix and the number of taxa, and that the Student's critical t values (t_{crit}) will therefore be very high for small data sets. In this way, the taxon sample size plays a known and constructive role in the probability estimates of a phylogenetic, cladistic examination.

If the relationship between E_{ij} and RAS_{ij} is curvilinear, natural log values can be used. In such instances, it is interesting to note that the null values for the natural log of E_{ij} and RAS_{ij} as determined by RER (eqs. 2 and 3) always result in a line without error and a slope (β_{null}) of 1.0. This property of RER suggests that the procedure results in a natural determination of the appropriate values of RAS_{null} and E_{null} (and therefore of β_{null}), and that t_{RASA} is an unbiased test for the presence of phylogenetic signal.

The question of the presence or absence of phylogenetic signal in any matrix (regardless of size) can be determined in polynomial time (via RASA); thus, the problem is "well solved" (Garey and Johnson 1979). The equation describing the cost (in computational steps) of the algorithm used to determine the observed and null pairwise values needed to calculate t_{RASA} after input is:

$$\mathrm{Cost} = \left[k\left(\frac{N*(N-1)*(N-2)}{6}\right) + 4(N*(N-1))\right] \qquad (4)$$

where k = number of characters, and N = number of taxa in the matrix. As equation (4) is devoid of exponentiation, the cost is polynomial; RASA is therefore efficient. The first term on the right in equation (4) is the product of the number of characters and the number of triplets possible with N taxa. The second term is the cost of the RER step. No additional cost is imposed by additional character states.

RASA can be applied using a specific outgroup by constraining all outgroup character states as potentially inherited, and therefore pleisiomorphic (i.e., a "rooted" measure of signal). Software currently under development includes the options of lognormal regression and outgroup choice. Anyone who wishes to receive a copy of this software should contact J.L.-W. Critical tests of the method follow.

Materials and Methods

RASA was performed on published data from a known phylogeny (Hillis et al. 1991), and data resulting from simulated branching evolution to demonstrate the application and interpretation of the test. All simulations were conducted as follows: an ancestral set of characters was constructed with L characters, each with C states. A topology was specified, and each branch was assigned $B \times \mu \times L$ mutations, where B = the length of the branch, as determined by topology (fig. 3), and μ = a mutation rate. This model of evolutionary change has

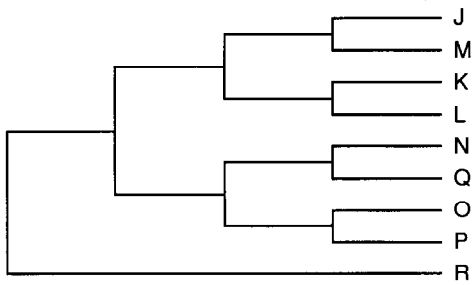

FIG. 1.—Topology used by Hillis et al. (1991) to conduct experimental phylogenesis using cultures of the bacteriophage T7. Taxon R represents the outgroup used.

the property of equal probability of mutation for all characters on a branch and for branches of the same length. Homoplasy occurs as a function of the number of character states for a particular character (C, B, μ). A set of iterations refers to 100 histories of evolutionary change for a given combination of C, μ, and L, and "evolved data" refers to taxon X character matrices that result from a set of iterations. Random data and mutations were generated with the random number generator in the Macintosh Toolbox.

Results

Correct Inference: Signal from a Known Phylogeny

Hillis et al.'s (1991) experimental phylogenesis of the bacteriophage T7 provides a good reference point to exemplify the application of RASA. Using the topology in figure 1, they subjected lineages of the T7 phage to a known mutagen, and retrieved 202 potentially informative restriction sites from the terminal and nodal populations. Every method of phylogenetic inference employed (Hillis et al. 1991) recovered the known topology; furthermore, some methods were extremely accurate at reconstructing ancestral character states (up to 98%).

Given the success of all phylogenetic methods tested in reconstructing the known history, this matrix should contain phylogenetic signal. The matrix of terminal lineages (excluding characters with missing values), when subjected to RASA, shows significant difference between the observed and null slopes of character covariation ($\beta_{obs} = 4.0235$, $\beta_{null} = 1.694$, $t_{RASA} = 11.316$, df = 24, $P < 0.001$; fig. 2*A*). If the appropriate outgroup is defined (taxon R, fig. 1), and outgroup character states are constrained, the test is slightly more significant ($\beta_{obs} = 4.521$, $\beta_{null} = 3.192$, $t_{RASA} = 12.684$, df = 17, $P < 0.001$; fig. 2*B*).

Precision and the Power of the Test

If t_{RASA} is an unbiased and reliable measure of the presence of phylogenetic signal with power, it should result in the correct inference both in the presence and absence of signal. First, if RASA is sensitive to signal, then when signal is present, t_{RASA} should be higher than the critical value (t_{crit}) for the specified degrees of freedom. Data matrices resulting from simulated branching evolution should contain phylogenetic signal; however,

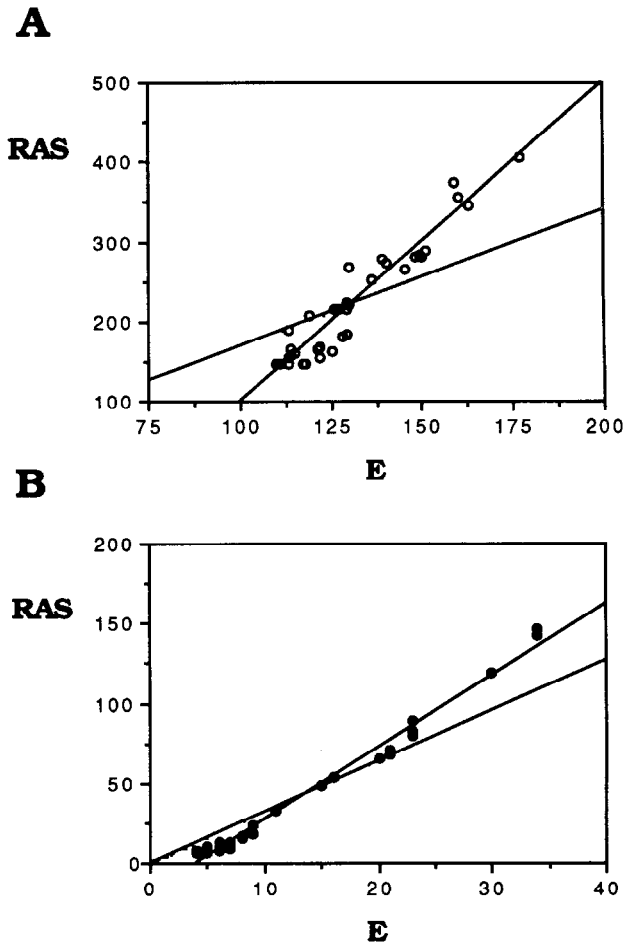

FIG. 2.—RASA plots (*A*, unrooted; *B*, rooted with taxon R) from the restriction site data from terminal populations of the bacteriophage T7 J–P in Hillis et al. (1991). Each point represents a taxon pair; *E* is a phenetic measure (number of informative character states shared by each taxon pair), and RAS is a measure of how unique the cladistic similarity is for each pair as determined by equation (1).

the amount of signal should vary with the following parameters: numbers of characters (L), number of alternative character states per character (C), mutation rate (μ), and branch length (B). If RASA has statistical power, then when signal is present, t_{RASA} should be greater than t_{crit}. The proportion of observations made where $t_{RASA} > t_{crit}$ is called power; 1 − power = probability of making a Type II error (failing to reject a false null).

A cursory test of the power of RASA was conducted using the seven-taxon topology in figure 3. The

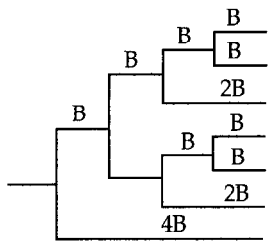

FIG. 3.—The seven-taxon phylogeny used for all simulations in this paper. *B* is the unit branch length, and all branch lengths are held constant to topology. All analyses are unrooted.

effect of saturation can be measured by manipulating μ, while the influence of the numbers of characters can be measured by manipulating *L*. *B* is held constant to the topology.

General Predictions

Prediction G1. The test statistic t_{RASA} should, in general, be positive for evolved data, but this is not to be expected for random data. t_{RASA} should be higher for evolved data than for random data, which should have a central tendency of 0.

C—*Alternative States per Character*

An increase in the number of character states should, in general, decrease the amount of convergence (homoplasy) under most models of evolution. Characters with greater numbers of alternative states (where transformation probabilities between states are the same) should be generally more informative on the basis that the saturation rate should be much slower with multistate characters. In general, multistate characters (evolving under this model) should retain more signal than binary state characters.

Prediction C1. The values of t_{RASA} should generally increase with an increase in *C* for evolved data, but should decrease with an increase in *C* for random data. Larger numbers of alternative character states should work to decrease the number of possible (random) intertaxon associations. Increasing the number of alternative character states should allow a greater proportion of mutations to remain informative (as homoplasy is less probable), thus constraining the associations among taxa to those reflecting history for evolved data.

Sets of iterations for 50 characters were conducted using evolved (μ = 0.04) and random data to test predictions G1 and C1. One hundred iterations were performed for each level of *C*. The overall results strongly favor the assertion that t_{RASA} tracks signal and noise and can discriminate between evolved and random data (fig. 4*A* and *B*, respectively). Both predictions G1 and C1 were satisfied (fig. 4); RASA therefore tracks signal as influenced by the number of character states. One unexplained observation is a slight positive central tendency in t_{RASA} with random data (fig. 4*B*). Use of overall similarity measures instead of *E* causes a dramatic loss rather than increase in power (unpublished data), so this not due to how *E* (informative similarity) differs from an overall measure of phenetic similarity. It is also noteworthy that the random data did not contain a single value higher than the t_{crit} for df = ((7 × 6)/2) − 7 − 2; the significance of the test is therefore high (<0.0001) under these conditions.

μ—*Mutation Rate*

Saturation is a likely condition for loci in long-separated lineages, and results in the attraction of long branches (Felsenstein 1978). Saturation will also occur at a given locus if it undergoes a high rate of mutation. In either case, informative differences between taxa will be lost as signal is, in effect, "erased" by multiple hits

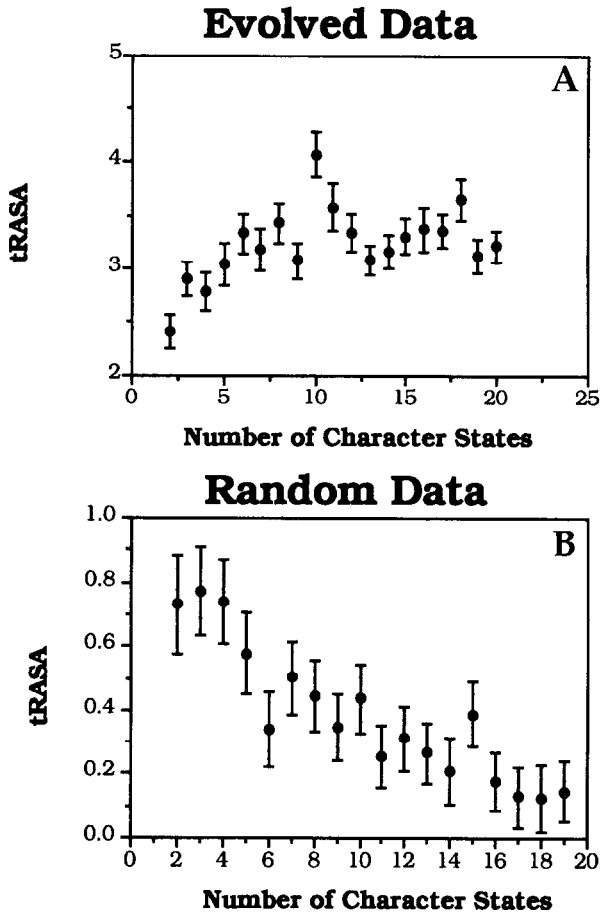

FIG. 4.—Mean t_{RASA} scores for 100 evolved (*A*) and random (*B*) matrices of seven taxa with 50 characters under a mutation rate of 0.04 × *B*. Closed points represent mean t_{RASA} values for each set of iterations (*L* = 50; μ = 0.04; 100 iterations of for each level of *C*).

at a given locus. Holding *B* constant to topology and varying μ is therefore equivalent to holding μ constant and varying *B*. Two predictions follow.

Prediction m1. Signal as measured by RASA should vary with increasing μ Trivially, signal should be zero if μ = 0; however, it should then increase to some intermediate level of μ, and then decrease again at high levels of μ as saturation begins to erode signal.

Prediction m2. At the point of complete saturation, the test statistic t_{RASA} from evolved data should approach t_{RASA} for random data (with equal *L* and *C*).

Results from sets of iterations for levels of μ = 0.04, 0.08, 0.16, 0.32, 0.64, 1.28, 2.56, and 5.12 were compared to test these predictions. (*L* = 50, *C* = 4). The curved relationship between μ and t_{RASA} was found, satisfying prediction m1 (fig. 5). The peak mean value of t_{RASA} was found at μ = 0.08 (fig. 5). Prediction m2 is also satisfied as mean t_{RASA} for μ = 1.28, 2.56, and 5.12 cannot be distinguished from t_{RASA} for random data (*L* = 50, *C* = 4).

L—*Number of Characters*

Data sets with few characters can be used to infer phylogeny (i.e., they can result in a tree); however, larg-

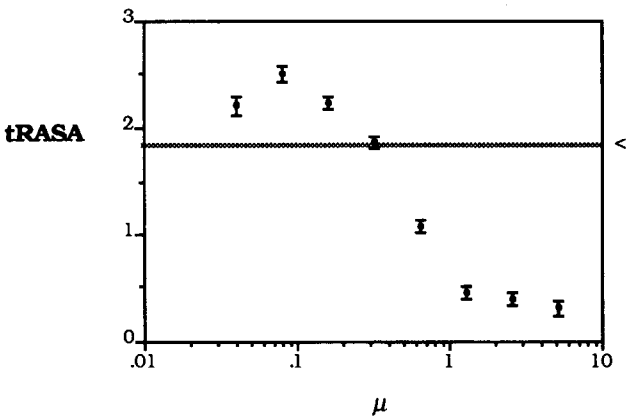

FIG. 5.—Low and high values of mutation (μ) on average result in submaximal levels of signal (t_{RASA}). Highest mean t_{RASA} values were found at intermediate levels. Saturation at high levels of μ is responsible for the reduction of signal. Evolved data only (L = 50; C = 4; 100 iterations each level of μ). Grey line indicates the critical Student's t-test value for 21 taxon pairs.

er data sets (higher L) will contain greater numbers of informative characters, more compatible characters, and therefore increased signal, and therefore more reliable estimates of phylogenies.

Prediction L1. Signal as measured by RASA should be found, on average, to increase with an increase in L with evolved data. This is expected because sets of compatible characters are less likely to be found with fewer characters.

Prediction L2. Variance in t_{RASA} should decrease with greater numbers of characters in evolved data because larger sets of compatible characters will more precisely record the signal : noise ratio.

To test predictions L1 and L2, sets of iterations at various levels of L (100 at each level ranging from 10 to 300; C = 4, μ = 0.04) were conducted. Random matrices were also generated for each level of L. Predictions L1 and L2 were found to hold. Mean t_{RASA} for evolved data was found to increase with greater numbers of characters (fig. 6*A*), while the variance (reported here as the standard error) decreases (fig. 6*B*). The increase in signal for evolved data is asymptotic (fig. 6*A*), reflecting the increasing precision with which t_{RASA} measures the amount of signal for the prescribed mutation rate, number of character states, and topology. It would be revealing to determine for a range of values for these parameters how many informative characters are required to reach this value.

Discussion

RASA is a new approach to the detection of phylogenetic signal grounded in statistical principle. RASA is compatible with all currently used tree-building algorithms. However, RASA also represents a significant philosophical departure from many existing methods in the field of phylogenetic systematics. These differences may cause some misunderstanding of RASA on several grounds. Of course, this method, like any statistical method of inference, also has its limitations. The following issues might be raised concerning the use of RASA:

1. The effect (however slight) of the positive central tendency in t_{RASA} with random data (fig. 4*B*) indicates a slight susceptibility to Type II error. This problem is strongly exacerbated by the use of absolute similarity instead of informative similarity (personal observation), suggesting that this source of error is minimized by using informative similarity (*E*). Nevertheless, the positive central tendency appears to be extremely slight.
2. It is possible that data sets with few taxa will appear to be devoid of signal due to small taxon sample size. Whether this is a limitation of RASA or a strength may be a matter of conjecture; a conservative view would call for higher stringency with fewer taxa to reduce the proportion of Type II errors. The inclusion of additional taxa in a study should alleviate the effect. We are currently investigating the influences of data set size on t_{RASA}.
3. Some data sets may include more than one set of covarying characters that reflects signal from conflicting tree topologies (e.g., phylogenetic history and convergent adaptation to similar environments). Conflicting but nonetheless real signal in such a matrix may mislead some measures of signal, or it may work to obscure the signal, depending on how it is measured. This may be a limitation of RASA in that it would fail to report that multiple contrasting signals do in fact exist; however, because it would indicate a lower amount of signal if internal heterogeneity exists in a given data set, its use would then prevent the construction (and application) of a potentially spurious phylogeny that reveals little about the histories recorded by different subsets of characters.

Research on the comparison of RASA to the *g1* moment (Hillis 1991) and the PTP tests (Faith 1991; Faith and Cranston 1992) for signal is needed. RASA differs from *g1* and PTP in that (1) it is an a priori measure of phylogenetic signal, (2) it is a tree-independent statistical measure of phylogenetic signal, and (3) it is not at all based on the assumptions of maximum parsimony. Methods that seek probabilistic support for a topology (or pattern) using that topology (or pattern) are inductive. For example, the bootstrap requires the assumption that the subsampling procedure provides a reasonable estimate of the variation that would be found in the character state universe if it were better sampled instead (Sanderson 1995). Because this assumption is untested, a bootstrap value can only provide a measure of the probability of inferring a node with the data at hand, rather than a measure of probabilistic support for the external validity of the resulting inference (i.e., generalizability is assumed, not tested by an independent criterion). No probabilistic support exists for inductive conclusions (Popper 1985). RASA is not inductive; it provides probabilistic support for the inference of the presence (or absence) of signal in the data in question. These differences mean that RASA may not share the

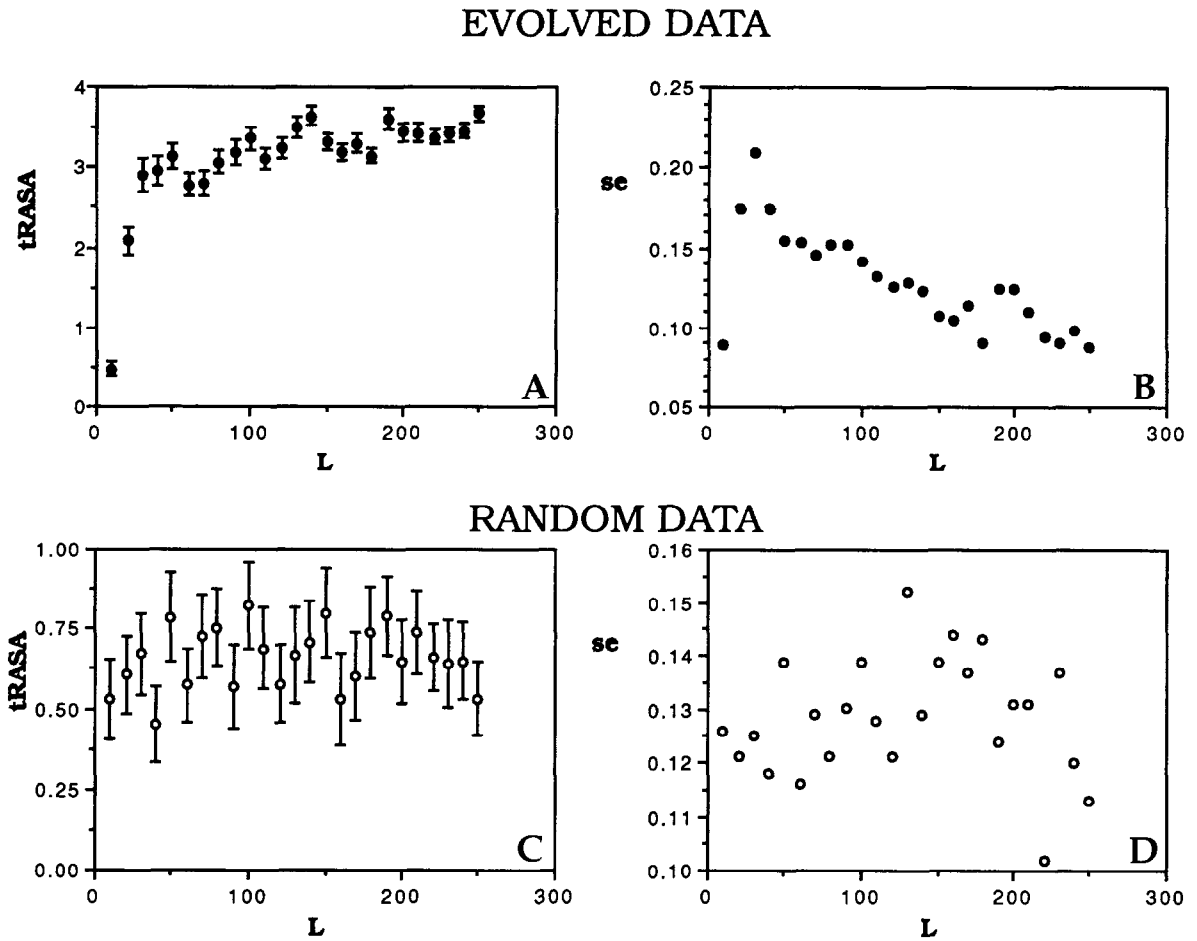

FIG. 6.—Data with few characters (L) are expected to result in low signal for both evolved and random data matrices (A and C, respectively), while deviation (here as the standard error) in t_{RASA} should decrease for evolved data (B). No trend in mean or variance was found for random data (C and D) (C = 4; μ = 0.04; 100 iterations each level of L).

limitations of $g1$ and PTP, and that the caveat may be withdrawn at some future point in time.

Maximum-likelihood methods (e.g., Felsenstein 1981) and character-weighting schemes are dependent on very specific underlying models of evolution, and require the application of general observations to specific instances. Although these approaches may be seen as exercises in the amplification of weakened signal, errors in the assumptions of such models can mislead the phylogenetic estimate (Cracraft and Helm-Bychowski 1991). Extrapolation of results from other studies that support particular models is inductive and a potential source of error in phylogenetic analyses. Even if the presumed processes are true in specific instances, extrapolation to specific cases may be perilous. Although such methods could be applied to the data prior to the measurement of signal with RASA, a principle strength of RASA is that it avoids inductive reasoning; therefore, we believe the use of such methods is philosophically inconsistent with the goals of RASA.

Another direction suggested by this study is the search for the general influence and interactions of the parameters used in this paper (C, L, μ, and a fourth parameter, rate variance between lineages) on the strength of phylogenetic signal left from a given history. Preliminary results indicate that signal as measured by RASA decreases under conditions with long branches (long edge attraction; Hendy and Penny 1989). Moreover, our ongoing work indicates that outgroup convergence does not seem to bias the test as it can for PTP (Trueman 1995) due to the fact that rooted RASA analyses constrain outgroup character states. It also appears that RASA does not conflate homoplasy with phylogenetic pattern unless concerted homoplasy greatly outweighs phylogenetic structure in the data. At this time, these assertions require further evaluation.

Finally, given the relative efficiency of the RASA algorithm, it would seem prudent to search for a tree-building algorithm based on the RASA approach, and find noninductive applications toward the identification of sets of covarying characters within a data set (cliques). We are currently working on each of these extensions.

Acknowledgments

Many people deserve credit for insight and criticism that have helped this approach mature. Sincere ap-

preciation is extended to Massimo Pigliucci, Carl Schlichting, Kent Holsinger, Michael Sanderson, and Joseph Felsenstein. Andrew Martin pointed to the need for a correction to the degrees of freedom. The comments of two anonymous reviewers were especially helpful. The prompt and thorough review process fostered by corresponding editor Simon Easteal and editor Barry Hall is especially appreciated. The appearance of a name in these acknowledgments should not be conflated with endorsement of the approach.

APPENDIX

The RASA approach requires careful consideration of the proper determination of the degrees of freedom for the homogeneity-of-slopes test. The issues are (1) phylogenetic nonindependence, (2) nonindependence due to redundant representation of each singular taxon $N - 1$ times in the regression, and (3) the number of terms that are estimated.

Issues 1 and 3 are straightforward: as phylogenetic nonindependence is what is being measured, it does not play a role, and two slopes and one intercept are estimated (the null intercept is always zero). Issue 3 is best approached from the beginning of the determination of RAS_{ij} and E_{ij}, and RAS_{null} and E_{null}.

On the surface, it would appear that in every case the number of degrees of freedom are $(N \times (N - 1))/2 - N - 3$. However, due to the redundant occurrence of taxa (e.g., taxon 1 is found in all pairs containing taxon 1), it may be considered that the appropriate degrees of freedom are at most $N - 3$. However, two pairwise matrices (containing the values of RAS_{ij} and E_{ij}) are used in this test. Both contain $(N \times (N - 1))/2$ entries. Both matrices are used to determine both nulls; i.e., RAS_{ij} and E_{ij} are used in the determination of RAS_{null} and E_{null}. Therefore, one matrix is not free to vary. Therefore, the appropriate degrees of freedom are

$N^2 - 2N - (N \times (N - 1)/2) - 3.$
{A − B − C − D}
A: Total number of cells in both matrices.
B: two diagonals.
C: one entire matrix.
D: two slopes and one intercept.

As it happens, $N^2 - 2N - (N \times (N - 1)/2) - 3 = (N \times (N - 1))/2 - N - 3$. This means that the number degrees of freedom can be most simply expressed as the number of pairs − the number of taxa − 3.

LITERATURE CITED

ARCHIE, J. W. 1989. A randomization test for phylogenetic information in systematic data. Syst. Zool. **38**:219–252.

CRACRAFT, J., and K. HELM-BYCHOWSKI. 1991. Parsimony and phylogenetic inference using DNA sequences: some methodological strategies. Pp. 184–220 *in* M. M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

CZELUSNIAK, J., and M. GOODMAN. 1995. Hominoid phylogeny estimated by model selection using goodness of fit significance tests. Mol. Phylogenet. Evol. **4**:283–290.

FAITH, D. P. 1989. Homoplasy as pattern: multivariate analysis of morphological convergence in Anseriformes. Cladistics **5**:235–258.

———. 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. **40**:366–375.

FAITH, D. P., and P. S. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. Cladistics **7**:1–28.

———. 1992. Probability, parsimony, and Popper. Syst. Biol. **41**:252–257.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

———. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **18**:368–376.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

FOREY, P. L., C. J. HUMPHRIES, I. L. KITCHING, R. W. SCOTLAND, D. J. SIEBERT, and D. M. WILLIAMS. 1992. Cladistics: a practical course in systematics. The Systematics Association Publication No. 10, Oxford Science Publications, Clarendon Press, Oxford.

GAREY, M., and D. S. JOHNSON. 1979. Computers and intractability: a guide to the theory of NP-completeness. W. H. Freeman and Co., New York.

GOLOBOFF, P. A. 1991. Homoplasy and the choice among cladograms. Cladistics **7**:215–232.

HENDY, M. D., and D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297–309.

HENNIG, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana.

HILLIS, D. M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. Pp. 278–294 *in* M. M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGETT, and I. J. MOLINEUX. 1991. Experimental phylogenetics: generation of a known phylogeny. Science **255**:589–591.

HILLIS, D. M., J. P. HUELSENBECK, and D. L. SWOFFORD. 1994. Hobgoblin of phylogenetics? Nature **369**:363–364.

HUELSENBECK, J. P. 1991. Tree length distribution skewness: an indicator of phylogenetic information. J. Hered. **88**:189–195.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum liklihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29**:170–179.

LOSOS, J. B., and F. ALDER. 1995. Stumped by trees? a generalized null model for patterns of organismal diversity. Am. Nat. **15**:329–342.

MADDISON, D. R. 1991. The discovery and importance of multiple islands of most parsimonious trees. Syst. Zool. **40**: 315–328.

MOOERS, A. Ø. 1995. Tree balance and tree completeness. Evolution **49**:379–384.

MYERS, R. H. 1990. Classical and modern regression with applications, 2nd edition. Duxbury Press, Wadsworth Publishing Company, Belmont, Calif.

NELSON, G., and N. PLATNICK, 1981. Systematics and biogeography: cladistics and vicariance. Columbia University Press, New York.

PENNY, D. E., E. WATSON, R. E. HICKSON, and P. J. LOCKHART. 1993. Some recent progress with methods for evolutionary trees. N. Z. J. Bot. **31**:275–288.

POPPER, K. R. 1985. The non-existence of probabilistic inductive support. Pp. 31–41 *in* I. MAHALINGAM and B. CARR, eds. Logical foundations: essays in honor of D.J. O'Connor. St. Martin's Press, New York.

SANDERSON, M. J. 1995. Objections to bootstrapping phylogenies: a critique. Syst. Biol. **44**:299–320.

SIMBERLOFF, D. 1987. Calculating the probabilities that cladograms match: a method of biogeographic inference. Syst. Zool. **36**:175–195.

STEEL, M. P. J. LOCKHART, and D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature **364**:440–442.

———. 1995. A frequency-dependent significance test for parsimony. Mol. Phylogenet. Evol. **4**:64–71.

TAUSCH, R. J., D. A. CHARLET, D. A. WEIXELMAN, and D. C. ZAMUDIO. 1996. Patterns of ordination and classification instability from changes in input data order. J. Veg. Sci. **6**: 897–902.

TRUEMAN, J. W. H. 1995. Sponges, plants and T-PTP. Syst. Biol. **44**:429–434.

WADDELL, P. J., D. PENNY, M. D. HENNY, G. ARNOLD. 1994. The Sampling distributions and covariance matrix of phylogenetic spectra. Mol. Biol. Evol. **11**:630–642.

WARNOW, T. J. 1993. Constructing phylogenetic trees using compatibility criteria. N. Z. J. Bot. **31**:239–248.

WILEY, E. O. 1981. Phylogenetics: the theory and practice of phylogenetic systematics. John Wiley and Sons, New York.

ZHARKIKH, A., and W.-H. LI. 1995. Estimation and confidence in phylogeny: the complete-and-partial bootstrap. Mol. Phylogenet. Evol. **4**:44–63.

SIMON EASTEAL, reviewing editor

Accepted February 14, 1996