Type I Error and the Power of the s-Test: Old Lessons from a New, Analytically Justified Statistical Test for Phylogenies

MARCOS A. ANTEZANA1 AND RICHARD R. HUDSON
Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717, USA; E-mail: marcos@uchicago.edu (M.A.A.), r-hudson@uchicago.edu (R.R.H.)

Abstract. — We present a new procedure for assessing the statistical significance of the most likely unrooted dichotomous topology inferrable from four DNA sequences. The procedure calculates directly a P-value for the support given to this topology by the informative sites congruent with it, assuming the most likely star topology as the null hypothesis. Informative sites are crucial in the determination of the maximum likelihood dichotomous topology and are therefore an obvious target for a statistical test of phylogenies. Our P-value is the probability of producing through parallel substitutions on the branches of the star topology at least as much support as that given to the maximum likelihood dichotomous topology by the aforementioned informative sites, for any of the three possible dichotomous topologies. The degree of statistical significance is simply the complement of this P-value. Ours is therefore an a posteriori testing approach, in which no dichotomous topology is specified in advance. We implement the test for the case in which all sites behave identically and the substitution model has a single parameter. Under these conditions, the P-value can be easily calculated on the basis of the probabilities of change on the branches of the most likely star topology, because under these assumptions, each site can become informative independently from every other site; accordingly, the total number of informative sites of each kind is binomially distributed. We explore the test’s type I error by applying it to data produced in star topologies having all branches equally long, or having two short and two long branches, and various degrees of homoplasy. The test is conservative but we demonstrate, by means of a discreteness correction and progressively assumption-free calculations of the P-values, that (1) the conservativeness is mostly due to the discrete nature of informative sites and (2) the P-values calculated empirically are moreover mostly quite accurate in absolute terms. Applying the test to data produced in dichotomous topologies with increasing internal branch length shows that, despite the test’s “conservativeness,” its power is much higher than that of the bootstrap, especially when the relevant informative sites are few.

When Felsenstein introduced the bootstrap method to phylogeneticists (Felsenstein, 1985a), he pointed out that the bootstrap requires three informative sites to produce a 95% significance in homoplasy-free data. He cautioned, however, that when the derived characters are less numerous, one nevertheless can have faith in the inference supported by them, if there is reason to believe that the characters are unlikely to be spurious. In other words, he acknowledged that the bootstrap does not necessarily give an accurate significance when the number of informative sites is small, and he realized that the probability of a given number of informative sites being spurious can somehow be assessed by considering other features of the data. Felsenstein also mentioned Templeton's pointing out to him that the requirement of three uncontradicted informative sites stood in contrast with Felsenstein's own clock-based test (Felsenstein, 1985b), which requires four such sites to give 95% significance. The contradiction between the number of uncontradicted informative sites required by each test to declare significance is restricted to cases with small numbers of competing informative sites (Fig. 1). Li and Zharkikh (1994) have in fact pointed out that the bootstrap...
procedure makes sense only when large numbers of competing informative sites are present in the data. Nevertheless, the discrepancy is an indication that the significance attributed to the signal by the bootstrap and Felsenstein’s clock-based test cannot both be correct. In fact, in its most general sense, the statistical significance of an inference is the complement of the probability that the support for the inference, or a more extreme one, was produced by processes other than the one whose significance is being tested. Such processes are finite and therefore this probability must have only one value.

A “microscopically justified” (sensu Maynard Smith, 1989:19) calculation of the P-value should not suffer from the above problems. Spurious phylogenetic signal (homoplasy) arise through parallel changes; therefore, a test that calculates explicitly the probability of such a pattern arising through parallel changes is likely to be more correct than others that make less-realistic assumptions. Methods like the bootstrap and Felsenstein’s clock-based test assume worst-case scenarios and do not exploit all the information available in the data for calculating P-values. As often stressed by proponents of the maximum likelihood method of phylogenetic inference, sequence data contain information not only about informative sites but also about the number of changes on the branches, the substitution model, etc. (Felsenstein, 1981). If one can determine with statistical justification a few crucial features of the evolutionary events that made the sequences in a data set different, then in principle, one can calculate the probability under an appropriate null model that such events produce (via parallel base changes) spurious support for the maximum likelihood dichotomous topology inferred from a data set. Informative sites are crucial in determining the maximum likelihood dichotomous topology and are therefore the most obvious target for a statistical test of phylogenies. This approach would therefore require (1) choosing a reasonable null hypothesis in which spurious informative sites can arise through parallel changes, (2) assuming an appropriate evolutionary model to be able to estimate accurately the probabilities of change per site under the null hypothesis and, from these, the probabilities that such changes create given numbers of spurious informative sites, and (3) adding up the probabilities of all cases in which any maximum likelihood dichotomous topology is supported by a number of informative sites that is at least equally unlikely as the number that supports the maximum likelihood dichotomous topology. We have therefore chosen to design an a posteriori test because a posteriori testing is the most common approach in molecular phylogenetics. Here, we discuss how to estimate such microscopically justified P-values for the case of four sequences when the Jukes and Cantor (1969) evolutionary model applies, and we use Monte Carlo simulations to study their performance in terms of type I error and power. We have called the general procedure the s-test because it is focused on the signal that supports the topological inference being tested.

**METHODS**

**Null Hypothesis**

For the null hypothesis, we have chosen the star topology with branch lengths estimated through maximum likelihood. The star topology rarely has a higher likelihood than any of the three dichotomous resolutions (Yang et al., 1995). However, such a topology can be shown to be most prone to produce spurious support for any of the three dichotomous topologies, relative to any of the three dichotomous topologies with branch lengths estimated by maximum likelihood. Estimating the branch lengths of the star topology with accuracy is crucial in implementing the s-test. Obviously, one could choose a star topology with branch lengths longer than the maximum likelihood ones and so make parallel changes more likely. We chose to use maximum likelihood to estimate the branch lengths of the star topology, because Monte Carlo simulations showed that this procedure estimates their average and variance accurately (Appendix 1, but see also below).
Support for Alternative Trees

Number of Informative Sites Supporting Best Tree

Figure 1. The significance degree obtained via the bootstrap (triangles) and Felsenstein’s clock-based test (circles) in the four-sequence case, as a function of the number of informative sites supporting the best topology and of the support for the other two alternative topologies. For instance, the points labeled (8,3,3) are the significances obtained with a data set in which the best-supported topology is favored by eight informative sites and the other two topologies are each favored by three informative sites. Open triangles indicate a bootstrap implementation under which when \( n \) best trees are inferred from a bootstrap pseudo-replication, each tree is considered supported \( 1/n \) times in that pseudo-replication (as done in PHYLIP parsimony bootstrapping). Solid triangles indicate a bootstrap implementation under which no tree is considered supported if two or more trees are best trees in the same pseudo-replication (this approximates the bootstrap with the Neighbor-Joining method).

Substitution Model

For simplicity, we have chosen the Jukes and Cantor (JC) evolutionary model (Jukes and Cantor, 1969), i.e., a model in which all sites are assumed to evolve identically and independently from each other, and every base is assumed to be equally frequent and equally accessible through substitution from any other base.

The P-Value

Assume that the data contain \( s \) informative sites congruent with the maximum likelihood dichotomous topology. Since we are interested in an a posteriori assessment of significance, we need to calculate the probability, under the null hypothesis, of observing spurious signal that supports any maximum likelihood dichotomous topology at least as strongly as the signal \( s \) supports the dichotomous topology originally inferred from the data. In fact, one is asking a question of the kind: What is the probability of getting, say, a five or higher on any die, when throwing three dice? Therefore, we need to conduct a “three-tailed” test.
We first need to calculate $P(S \geq s)$, the probability that the random variable $S$, the number of spurious informative sites congruent with the original dichotomous topology, assumes a value equal to $s$ or higher. If the JC model applies, an informative site congruent with the original dichotomous topology can arise at each site, or not, independently from what happens at other sites, and therefore the number of such informative patterns over $L$ sites is binomially distributed with mean $LH_i$ and variance $LH_i(1 - H_i)$, where $H_i$ is the probability per site that such an informative site arises. $P(S > s)$ is therefore a simple but long sum of binomial probabilities; its complement

$$P(S > s) = \sum_{i=0}^{s-1} \binom{L}{i} H_i^i(1 - H_i)^{L-i}$$

is usually shorter. The calculation of $H_i$ is also straightforward: Consider a star topology with four branches and assume that a given base present at a given site of the sequence at the sole internal node of the star topology, is different from the bases at the same site in the terminal node sequences A, B, C, and D with probabilities $a$, $b$, $c$, and $d$, respectively. Assume that the maximum likelihood dichotomous topology separates A and B from C and D. An informative site that supports this topology can arise through either two or four parallel changes: (1) Two changes to the same base could occur on the branches going to A and B at a given site, while no changes occur at that site on the branches going to C and D; or (2) the two changes occur on the branches going to C and D, and no changes occur on those going to A and B; or (3) four changes occur at the site, two leading to a second base in A and B and the other two leading to a third base in C and D. Therefore $H_i$ is equal to

$$H_i = \frac{1}{3} [ab(1 - c)(1 - d) + cd(1 - a)(1 - b)] + \frac{2}{27}abcd$$

(2)

The probabilities $H_2$ and $H_3$ for the production of informative sites that support the other two dichotomous resolutions can be obtained by swapping in appropriate ways $a$, $b$, $c$, and $d$.

$P(S \geq s)$ is the probability of spurious support for the originally inferred dichotomous topology, but the fact that similar support could have arisen for any of the other two resolutions as well needs to be considered. In the dice example introduced above, “three-tailing” is simple because the dice are independent from each other and identical to each other. In the star topology, the total number of spurious informative sites of each kind can be considered independent from the number of the other two kinds only when the total number of informative sites is well below the order of magnitude of the sequence length, i.e., when the three numbers behave like independent Poisson variables. In our simulations the maximum probability of change per site will be 0.3, which in equation 2 gives a maximum value for $H$ of 0.03, when all branches are equal. Such an $H$ results in only 9% of all sites being informative on average, in any of the three directions; thus, it appears safe to assume that the support for each of the three topologies is independent from the support for the others.

However, unlike in the dice example, in the star topology the probability of producing informative sites for each of the three topologies is almost always different. In other words, our situation is equivalent to throwing three dice of which each has a different number of sides, say, 6, 8, and 10. Obviously the probability of producing a five or higher is much larger for the 10-sided die than for the other two dice.

The exact formula for “three-tailing,” i.e., for calculating the probability that none of the three topologies is as strongly supported as the maximum likelihood dichotomous topology, must therefore be

$$1 - [1 - P(S \geq s)][1 - P(T \geq t^*)][1 - P(U \geq u^*)]$$

(3)

where $s$ is still the number of observed informative sites supporting the maximum likelihood dichotomous topology, $T$ and $U$ are the corresponding random variables for the numbers of informative sites congruent with the other two dichotomous topologies, and $t^*$ and $u^*$ are the numbers of informative sites for which $P(T \geq t^*)$ and $P(U \geq u^*)$ are
most similar to, but not larger than, \( P(S \geq s) \). In plain words, the three-tailing formula in formula 3 calculates the probability of non-occurrence of supports that are equally or more unlikely than the support for the originally inferred maximum likelihood dichotomous topology as quantified by \( P(S \geq s) \), for any dichotomous topology.

The alternative formulation \( 1 - \left[ 1 - P(S \geq s) \right] \left[ 1 - P(T \geq s) \right] \left[ 1 - P(U \geq s) \right] \), is adequate—and thus equivalent to formula 3—only when the probability density functions of \( S \), \( T \), and \( U \) are identical or—to use again the dice comparison—when the three dice are identical. Otherwise, this formula can lead to errors. Assume for instance that the null hypothesis star topology has long-branch-attraction structure (Felsenstein, 1978). In such a situation, the informative sites that support the dichotomous topology in which the two terminal taxa at the end of the two long branches appear together are thus much more likely to arise than are sites supporting the other two topologies. Therefore, whenever the maximum likelihood dichotomous topology is not the topology favored by long-branch-attraction homoplasy, the \( P(T \geq s) \) or \( P(V \geq s) \) in this alternative formula becomes very large, because \( s \) or more sites supporting the long-branch-attraction topology are very likely to arise. This should lead to overestimates of the \( P \)-value and should thus compromise the power of the test. In other words, we need to calculate the probability of producing spurious signal that supports a maximum likelihood dichotomous topology at least as strongly as does the signal that supports the original dichotomous resolution. Therefore, less-extreme supports should be disregarded. Conversely, if the maximum likelihood dichotomous topology is the long-branch-attraction dichotomous topology, the \( P(T \geq s) \) or \( P(V \geq s) \) in this alternative formula become spuriously very large and severely depress the \( P \)-value, making the test non-conservative.

A simplification of formula 3 would be the formula

\[ 1 - [1 - P(S \geq s)]^3 \]  

which simply assumes that the same support as that observed could also have arisen for the other two topologies and independently so.

Although more complex, formula 3 should deliver higher statistical power; it exploits the fact that because of discreteness, the values of \( P(T \geq t^*) \) and \( P(U \geq u^*) \) can be substantially smaller than that of \( P(S \geq s) \) even in clock situations, where \( S \), \( T \), and \( U \) are supposed to be identical (see below). We will explore with simulations the type I error and power of the two approaches to obtain the final “three-tailed” \( P \)-value of the \( s \)-test formalized in formulae 3 and 4, referring to these two \( P \)-values as \( s \)-discr and \( s \)-cubed, respectively.

**Estimation of the Probabilities of Change**

The probabilities of change \( a \), \( b \), \( c \), and \( d \) on the branches of star topology were estimated by using the maximum likelihood routine \( dnaml \) in PHYLIP 3.5c (Felsenstein, 1993) and a star topology as the user-defined tree. From the branch lengths (\( u \) values) in the “treefile” produced by \( dnaml \), the corresponding probabilities of change were calculated according to the Jukes and Cantor (1969) formula

\[ P(\text{change}) = \frac{3}{4} \left( 1 - e^{-\frac{1}{4}wu} \right) \]  

**Simulations**

Each simulation of evolution on a given topology started with a new internal sequence generated by assuming that the probability of appearance of each base was 0.25. On every given branch, all sites of the sequence had the same probability of change.

**Simulations to study type I error.**—We studied the type I error of the \( s \)-test by applying it to sequence data produced by simulating the evolution of four DNA sequences of length \( L = 100, 1,000, \) and \( 10,000 \) bases in star topologies in which all branches had the same length (“molecular clock”) or in which two branches were short and the other two were long (long-branch-attraction). The probabilities of change on each of the four branches were 0.01, 0.03, 0.06, 0.1, or 0.3 in topologies with a clock, whereas in those with a long-branch-attraction structure, two
branches had probability of change of 0.01 and the other two each had probability 0.03, or 0.06, or 0.1, or 0.3. Type I error was estimated by using 10,000 simulations.

Simulations to study power.—To explore the power of the s-test, we examined its ability to reject the null hypothesis of the star topology as the probability of change along an internal branch that separates sequences A and B from C and D increases. We have done so for topologies with a clock and terminal branch probability of change of 0.03 or 0.3; for the long-branch-attraction topologies, we examined cases in which the branches going to B and C had probability of change of 0.03 or 0.3 and the shorter ones (to A and to D) had probability 0.01. Note that in the latter topologies, the internal branch creates one of the two dichotomous topologies that are not favored by long-branch-attraction homoplasy. We used sequences 300 and 1,000 bases long for both kinds of trees. We compared the power of the s-test to that of the bootstrap. For clock data, we wrote a parsimony-based bootstrap algorithm that does not add fractional support to any tree when a pseudo-replicated data set supports more than one tree (see also Fig. 1). Because unweighted parsimony fails with long-branch-attraction data, we wrote a bootstrap algorithm based on the Neighbor-Joining method (Saitou and Nei, 1987) and corrected distances with the JC formula. Power was estimated by using 10,000 simulations of each dichotomous tree examined.

RESULTS

Type I Error

Figures 2 and 3 show the cumulative distribution functions (c.d.f.) of s-discr and s-cubed obtained from simulations of star phylogenies. The c.d.f. curves indicate that the number of rejections at almost any given level of significance would be too few if one were to use the plain complements of s-discr and s-cubed as degree of significance, especially when dealing with short sequences and small probabilities of change. In other words, the two statistics appear to be often conservative because the type I error expected from using them when the null hypothesis is true is too low. As expected, the type I error of s-discr is higher than that of s-cubed.

We studied three possible reasons for the apparent conservativeness of the estimated s-statistics during the simulations. A first possible cause is simply the discrete nature of the tested observation. Indeed, a conservative c.d.f. is always expected when events are very discrete, even when P-values are calculated exactly: For instance, assume a process where 0, 1, 2, and 3 counts arise with probabilities 0.889, 0.1, 0.01, and 0.001, and we can calculate these probabilities exactly. If we simulate this process many times, we will observe that only in ~1.1% of the trials, not 5.0%, is the calculated probability ≤ 0.05. We could therefore conclude erroneously that our calculation of probabilities in the proximity of 0.05 is biased upwards. In the case of the s-test, the number of changes on the branches varies from trial to trial and thus the probability of having the same number of congruent informative sites varies from trial to trial as well. This explains why the c.d.f.'s in Figure 2 often appear smooth rather than steplike (the step-like c.d.f.'s of the cases with low probabilities of change and L = 100 are due to the discreteness of the realized number of changes on the branches, not to that of s, which in these cases is never > 1). A second factor that can make the s-test conservative is overestimation by maximum likelihood of the probabilities of change and L = 100 are due to the discreteness of the realized number of changes on the branches of the star topology in those trials in which the s count is large (remember that over all trials dnaml performs well; see Appendix 1). This would bias s-cubed towards higher values when s is large, but no prediction is possible for s-discr. A third possibility is that the assumption of the binomial in equation 1 may bias the s-statistic upwards. In fact, binomial probabilities can be calculated for any number of informative sites below L, but the number of changes that actually occur on the branches in a given simulated trial sets a much lower upper limit. Thus the binomial probabilities for numbers of informative sites > s but below the L limit must be overestimated, and this should inflate the real value of P(S ≥ s), P(T ≥ t*), and
FIGURE 2. The cumulative distribution function of s-discr (top plots) and s-cubed (bottom plots) in star topologies with molecular clock. On the vertical axis is the number of trials having a given $P$-value or lower. The straight (gray) line is the c.d.f. for a test with correct type I error (e.g., it would give 500 cases with $P < 0.05$ out of 10,000 simulated cases). Thin black lines, thicker gray lines, thicker black lines, thickest gray lines, and thickest black lines are cases in which the probability of change per site on every branch was 0.01, 0.03, 0.06, 0.1, and 0.3, respectively. From left to right, sequences were 100, 1,000, and 10,000 bases long.
FIGURE 3. The cumulative distribution function of s-discr and s-cubed (top and bottom plots, respectively) in star topologies with a long-branch-attraction structure. On the vertical axis is the number of trials having a given $P$-value or lower. The straight (gray) line is the c.d.f. for a test with correct type I error. Thinner gray lines, thinner black lines, thicker gray lines, and thicker black lines are cases in which the probability of change per site on the two longest branches was 0.03, 0.06, 0.1, and 0.3, respectively (on shortest branches it was 0.01). From left to right, sequences were 100, 1,000, and 10,000 bases long.
$P(U \geq u^*)$, $P(S \geq s)$ is indeed totally dependent (conditional) on the number of changes that occur on the branches and not directly on the probability that can take place. Transforming these changes into probabilities of change and using the binomial is useful in obtaining a description of the null model when the null model is not valid. When the null model is valid, however, the realized number of changes in a given trial is a real entity, i.e., a fixed quantity, not a probability; and if parallel changes occur in the trial, it is because some of these very changes change specific sites in parallel.

One could argue that because the type I error becomes more correct with longer sequences, discreteness must be the cause of the observed conservativeness. The estimation of branch lengths by `dnaml` clearly becomes more accurate as the sequences get longer, but we do not know whether the binomial assumption improves its performance as well. Therefore, we explored the relative importance of the above factors in biasing the two empirically estimated $s$-statistics. For each simulated trial we calculated two variant statistics called $s_r$-discr and $s_r$-cubed ($t$ stands for "true"), both of which are based on the true values of $P(S \geq s)$, $P(T \geq t^*)$, and $P(U \geq u^*)$. The latter set of probabilities were estimated by repeating 100,000 times the changes that occurred in the branches in the given trial, allowing them to occur at any site but not in other branches, and keeping track of the kind and number of the informative sites that arise (note that an analytical calculation of say $P(S \geq s)$ would involve nested hypergeometric random variables and would be very arduous if at all feasible). We also calculated a second variant of the $s$-statistics, keeping the multinomial assumption, but also calculating the probability of change for each branch by dividing $x$ by $y$ the length of the simulated sequence the realized number of changes that occurred on the branch in the given trial. We called these variants $s_r$-discr and $s_r$-cubed ($r$ stands for "realized"). Comparing the c.d.f.'s of the two empirically estimated $s$-statistics with those of these variants allowed us to characterize the causes of the conservativeness of the $s$-test.

Figure 4 shows the c.d.f.'s of the above statistics for a star topology with a clock, 100- and 1,000-base sequences, and probabilities of change of 0.01 and 0.1. Overall, the c.d.f.'s of all three kinds of statistics are quite similar, although the two $s_r$-statistics are the least conservative in all cases and, with a single exception, the two $s_r$-statistics are in turn less conservative than the empirically estimated ones. This indicates that inaccuracies in the maximum likelihood estimation of branch lengths as well as errors resulting from the binomial assumption are involved in making the empirically estimated statistics more conservative. The error from the binomial is most apparent when probabilities of change are large, because then the $s_r$- and the empirical statistics, which are both binomial-based, are almost overlapping—though clearly more conservative than the $s_r$-statistics. The bias due to the maximum likelihood estimation of the probabilities of change becomes evident when comparing the three versions of $s$-cubed in situations with small probabilities of change. Under such conditions $s_r$-cubed and $s_r$-cubed overlap, indicating that the binomial assumption does not cause much bias. However, these statistics are less conservative than the empirical ones, indicating that the maximum likelihood estimation of the probabilities of change is the cause of the additional conservativeness of the estimated statistic. Surprisingly, when the probability of change is 0.01, the c.d.f. of $s$-discr is less conservative than that of $s_r$-discr for $P \leq 0.05$. We do not know the cause of this but believe that, somehow, branch estimation errors affect the exploitation of discreteness by $s$-discr and thereby deflate the statistic. As we go from 100 to 1,000 bases, the match of the three kinds of statistics becomes closer, especially with low probabilities of change. This probably happens because probabilities of change are estimated more exactly from longer sequences. Similar trends were observed for long-branch-attraction star topologies and for other probabilities of change (not shown).

As is clear from Figure 4, however, removing the errors resulting from the estimation of branch lengths and from the binomial as-
The cumulative distribution function (c.d.f.) of \( s_{i} \)-discr, \( s_{i} \)-discr (top plots), \( s_{i} \)-cubed, \( s_{i} \)-cubed (bottom plots), and of the corresponding empirical \( s \)-statistics for star topologies with a clock, 100-base (left plots) and 1,000-base sequences (right plots), and probabilities of change on each branch of 0.01 or 0.1, respectively. The straight (gray) line is the c.d.f. for a test with correct type I error. The c.d.f.’s of \( s_{i} \), \( s_{i} \), and of the empirical \( s \)-statistics are shown as thicker black lines, thinner black lines, and thick gray lines, respectively. C.d.f.’s are based on 10,000 replications; the estimation of the \( s \)-statistics was based on 100,000 replications of the configuration of base changes that occurred in the corresponding trial.

Assumption does not make the type I error much more correct. This indicates that the conservativeness of the \( s \)-test must be due largely to discreteness.

We have, however, produced a more decisive proof that discreteness is the cause of the conservativeness of the \( s \)-tests, by applying a continuity correction to \( s_{i} \)-cubed. This correction is similar to that used to obtain a correct type I error when applying Fisher’s exact test to low count data. The procedure is as follows: For each given trial, one calculates the one-tailed \( P \)-values for \( s \) and for \( s + 1 \) informative sites; choose randomly a value between these two \( P \)-values, assuming that the value can fall anywhere between the two with equal probability; finally, one uses this third value—instead of \( P(S \geq s) \)—in formula 4 to get the adjusted \( s \)-cubed statistic, \( s_{i} \)-cubedadj. If \( P(S \geq s) \) and \( P(S \geq s + 1) \) are calculated correctly every time, then the c.d.f. of \( s_{i} \)-cubedadj will be a straight line with 10% type I error at the 90% level, 5% at the 95% level, 1% at 99%, and so forth.

In star topologies with a clock or a long-branch-attraction structure, and for the probabilities of change for which we studied the type I error of \( s \)-cubed, we found that the c.d.f.s of \( s_{i} \)-cubed adj were invariably correct (Fig. 5). Thus this exercise con-
Firmly confirmed that the conservativeness of $s_c$ is mostly due to discreteness. Moreover, applying the correction on $s_c$ also gives a conservative c.d.f., even in low-homoplasy situations. However, a hybrid approach, in which the upper $P$-value we used (for $s + 1$) was the true value of $P(S > s + 1)$, resulted in almost correct c.d.f.s except when probabilities of change were high (not shown). This indicates again that the binomial assumption is acceptably accurate for calculating $P(S = x)$ when $x$ is close to a trial's $s$ count, but it overestimates $P(S = x)$ for higher counts, thereby biasing the adjustment towards higher $P$-values. It is not possible to produce a similar proof for $s_{\text{discr}}$, however, because this statistic exploits discreteness to improve its power and thus cannot be made "continuous."

As pointed out above, the c.d.f.'s of $s_c$ and $s_{\text{discr}}$ are often very similar to those of the empirically estimated statistics, an indication that the latter statistics may be nearly exact. However, that two c.d.f. curves overlap closely is no proof for a point-by-point correspondence between the two statistics considered. To ascertain whether the estimated statistics are close to their true values, we plotted the difference between each estimated $s$- and $s_{\text{discr}}$-statistic and their $s_{\text{r}}$-versions (Fig. 6). This was done for star topology data with a clock, with $L = 100$ and 1,000 bases, and with probabilities of change equal to 0.01, 0.03, and 0.1. Figure 6 confirms that "errors" in the maximum likelihood estimation often cause marked discrepancies between the estimated statistics and the corresponding $s_r$-statistics since the $s_r$-statistics are often much closer to the $s_r$-statistics. Fortunately, in most cases the empirically estimated $s$-statistics are too high, which makes them conservative. Importantly, the plots also show that longer sequences and intermediate probabilities of change favor the estimation of the $s$-statistics. Also noteworthy is that $s_{\text{discr}}$ appears to be estimated less accurately than $s_c$. This decreased accuracy is possibly due to the fact that errors in branch length estimation and errors from the binomial assumption can affect $s_{\text{discr}}$ through all three of its terms, not only through one term, as in $s_c$; errors in $s_{\text{discr}}$ are therefore compounded (formulae 3, 4).

**Power**

Figure 7 shows plots of the power of $s_{\text{discr}}$, $s_c$, and the bootstrap when
FIGURE 6. On the vertical axis is the departure of the value of $s_{\text{discr}}$ (circles) and each empirically estimated $s_{\text{discr}}$ (minuses) from their corresponding true value, $s_{\text{discr}}$, for each simulated case (upper four plots), and the departure of each value of $s_{\text{cubed}}$ and $s_{\text{cubed}}$ from the corresponding $s_{\text{cubed}}$ (lower four plots). Each of these differences was plotted against the applicable $s_{\text{cubed}}$ value. Star topologies had clock structure and probabilities of change of 0.01 or 0.1. Sequences were 100 or 1,000 bases long.
the topology has equally long terminal branches. As the plots show, both s-tests have higher power than the bootstrap, especially in assigning higher-level significance. In cases with a clock and terminal branch probability of change equal to 0.3, the powers of the tests are more similar, although the s-tests are still substantially superior, especially for declaring significance at higher levels. In long-branch-attraction topologies, the s-tests were superior to the bootstrap in low homoplasy situations but, contrary to the clock cases, they had an even higher power when the two long branches had a probability of change of 0.3 (not shown). Notably, the large differences in type I error between s-discr and s-cubed do not seem to lead to substantial differences in power relative to the power of the bootstrap.

Conclusions

Type I Error

We showed above that the s-tests are conservative and proved that this is mostly due to the discrete character of informative sites. We also found that s-discr has a more correct type I error than s-cubed. However, comparisons of the estimated P-values with their "true" values indicate that s-discr is often less accurate than s-cubed. The fact that the discreteness correction of s-cubed resulted in a correct type I error indicates that informative sites arise more or less independently in the parameter range explored and that discreteness is the main cause of the conservativeness of the tests. The estimation by maximum likelihood of the probabilities of change in the star topology was found to be quite accurate for the purpose of calculation of s-cubed [but only for moderately long sequences (L \geq 1000 bases)], and so was the binomial assumption. However, the errors produced when calculating s-cubed in shorter sequences are almost always biased towards conservativeness, so that, as far as type I error is concerned, this test can be applied safely in such situations. This is not, however, true of s-discr, which is underestimated more often when L is small.

An important implication of the fact that discreteness can depress type I error markedly is that one should not determine significance cutoffs simply from the simulated c.d.f.s of statistics like the s-tests. Indeed, suppose that the c.d.f. of s-cubed could actually be used to determine the significance cutoffs for the case L = 100 and probability of change = 0.1. In that case, one would produce the "desired" 5% type I error at the 95% significance level by declaring "significant at the 95% level" all cases with s-cubed \leq 0.12 (see Fig. 2). This is patently misleading, given that in such cases the real P-value is, obviously, 0.12 and the significance is 88% rather than 95%. Simulation-determined significance cutoffs can be assayed for the presence of bias attributable to discreteness by applying a discreteness correction such as the one above, that is, by using the probabilities corresponding to the proposed significance cutoffs in the discreteness correction procedure and checking whether the procedure delivers a correct c.d.f.

Although the values of s-discr and s-cubed were calculated under almost ideal conditions such as minimal number of sequences, simple and known evolutionary model, and so forth, we nonetheless found them to be somewhat approximate. One can therefore question whether a statistical test for phylogenies based on terminal-node sequence data will ever be exact. Figure 6 shows, however, that the empirically estimated s-cubed can be very close to its real value when sequences become moderately long, especially in low-homoplasy situations. This length requirement is more moderate than the hundreds of informative sites that the bootstrap requires in order to be more accurate (Li and Zharkikh, 1994).

Power

We showed that both the s-cube and the s-discr test have a much greater power than the bootstrap in cases with low homoplasy (i.e., short terminal branches), both in topologies with terminal branches of equal length and in topologies with a long-branch-attraction structure. This is so because, in both these cases, the s counts required in the bootstrap procedure to produce high significance values are unlikely to arise when
the probability of change on the internal branch is low. For instance, the bootstrap cannot declare significant at the 95% level a dichotomous reconstruction supported by a single unopposed informative site; instead, it will always give a 64% value, or 75.5% if we allow for fractional support in the four-sequence case. The two $s$-tests, in contrast,
can declare such significance for any value of \( s > 0 \), if the relevant pairs of terminal branches are short enough (formula 2).

On the other hand, in cases with high homoplasy and a clock, the s-tests and the bootstrap have more comparable power. The reason for this similarity in power is that when homoplasy is high, that is, when the number of competing informative sites is large, the bootstrap and the s-test require similar s counts to declare significance. This indicates then that the differences in the number of informative sites supporting the competing topologies assayed by the bootstrap's pseudosampling exercise, reflect well the probabilities of the sites having arisen through parallel changes. This is also the reason why the type I error of the bootstrap in such situations (~2% at the 95% level in the four-sequence case; Zharkikh and Li, 1992a, 1992b; Li and Zharkikh, 1994) is comparable with that of the s-test. Another indication that similarities in type I error between the two testing approaches correlate well with similarities in power is the fact that the power of both tests at the 75% level is very similar in every situation we examined (not shown) and, moreover, that the type I error of the bootstrap at the 75% level in the four-sequence/high-homoplasy case is almost correct (Li and Zharkikh, 1994). On the other hand, in topologies with high long-branch-attraction homoplasy, the s-tests were much more powerful than the bootstrap. This contradiction is spurious, however, because an examination of the data shows that the s-tests' superiority in these cases again is related to the fact that the s-counts produced in these cases are too low for the bootstrap to declare a high significance.

In all power assays, the s-tests were superior to the bootstrap in assigning high significance values. This is because, unlike the bootstrap, the s-test can declare such significance also for lower s counts—not just when s is large.

Strikingly, the two s-tests do not differ by much in their power, despite the fact that they have quite different type I error in low-homoplasy situations. The explanation is the following: Their differences in type I error result from their different use of the information on terminal branch changes, but in the power assays the internal branch produces informative sites independently of what happens in terminal branches. Therefore, that most non-spurious informative sites tested in the power assays, are equally significant for both s-tests.

**Extension to More Complex Situations**

We implemented the s-tests under ideal conditions for the smallest number of sequences possible, because we wanted to assess the best performance this testing approach can deliver. The tests, however, will have to overcome implementation problems if they are to be used for topologies with more sequences or with more-complex evolutionary models. The first problem is to determine which kinds of base patterns at a site actually support a branch of interest and which patterns not actually present in the data nonetheless could have been present and could have supported the branch. Once all such patterns are known, we must be able to quantify their relative weight as supporting signal—both as single sites and, more importantly, in multiple presence in the company of heterogeneous combinations of other "informative" sites. We will also need then to evaluate the sum of the multinomial probabilities of all configurations that give equal or higher support than that originally observed, in a tree where the tested branch has been set to zero. All of these steps are potentially solvable, especially for not very complex models. For instance, Zharkikh and Li (1992b) have presented the equivalent of equation 2 for the two-parameter model.

**Why the s-Test and Not a Likelihood Ratio Test?**

The s-tests require estimating the most likely dichotomous tree and the most likely star tree. Thus one could ask, why not use a likelihood ratio test instead? The likelihood ratio test, in fact, should have all the advantages of the s-test and be simpler to implement because it requires a mere division of two likelihoods to obtain the ratio. However, the s-test delivers a \( P \)-value directly, whereas translating a likelihood ratio into a \( P \)-value requires knowing the null distribution of the ratio. Unfortunately, even with
sequences as long as 10,000 bases, the null distribution of the likelihood ratio in the four sequences case is not exactly χ²1 df over the range of probabilities of change studied here for the s-tests (Antezana and Hudson, in prep.). Thus, for reasonably long sequences, the distribution has to be estimated with computer-intensive simulations on a case-by-case basis. Furthermore, since even these simulated distributions must be “conservative” because of discreteness, as are those of the s-statistics, one cannot naively read the significance cutoffs directly from them. How to deal with this last problem is not known to us. Therefore, the choice is between implementing the s-test for more-complex situations, finding a way to predict analytically the null distribution of the likelihood ratio for data of any discreteness, and devising ways to obtain these distributions from fast simulations that in some way correct for the discreteness bias.

Program for Calculating s-discr and s-cubed

A Mathematica notebook is available by request from M.A.A. to calculate s-discr and s-cubed. It assumes the JC model and requires the user to enter s, a, b, c, and d.

Acknowledgments

We thank Steve Frank, Harry Mangalam, George Gutman, and Terry Gaasterland for computer time, Walter Fitch for encouragement, and Carlos Machado, Molly Przeworski, Joy Bergelson, Marty Kreitman, John Wakeley, and Michael Sanderson for comments and proofreading. Joseph Felsenstein kindly clarified issues related to dnaml. Special thanks goes to Michael Charleston for his insightful comments and to Martha Wayne and the NCSU reading group in statistics for provoking MAA by calling formula 4 “a conservative approximation”. MAA’s work was partially supported by a postdoctoral fellowship of the Schweizerischen National Fonds and by NIH grant no. GM39355 to Martin Kreitman.

References


Received 19 May 1998; accepted 15 October 1998
Associate Editor: M. Sanderson
### Appendix 1. Simulated and *dnaml*-estimated base changes per site in a star topology with four sequences.

<table>
<thead>
<tr>
<th>PC</th>
<th>Realized changes per site</th>
<th><em>dnaml</em>-estimated</th>
<th>Realized changes per site</th>
<th><em>dnaml</em>-estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Branch</td>
<td>100</td>
<td>1,000</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.949</td>
<td>0.992</td>
<td>1.025</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.826</td>
<td>1.031</td>
<td>1.131</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.030</td>
<td>0.030</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.593</td>
<td>0.517</td>
<td>0.598</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.342</td>
<td>0.204</td>
<td>0.373</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.060</td>
<td>0.060</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.387</td>
<td>0.376</td>
<td>0.392</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.108</td>
<td>0.197</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.064</td>
<td>0.108</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.096</td>
<td>-0.053</td>
<td>-0.024</td>
</tr>
<tr>
<td></td>
<td>Long-branch attraction</td>
<td>100</td>
<td>1,000</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.030</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.528</td>
<td>0.975</td>
<td>1.015</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.224</td>
<td>0.939</td>
<td>1.210</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.060</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.368</td>
<td>0.968</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.166</td>
<td>1.025</td>
<td>0.773</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.100</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.262</td>
<td>0.973</td>
<td>1.006</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>-0.018</td>
<td>0.942</td>
<td>1.053</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.300</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.002</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.048</td>
<td>0.998</td>
<td>1.003</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>0.018</td>
<td>0.995</td>
<td>0.850</td>
</tr>
</tbody>
</table>