Sampling Properties of the Bootstrap Support in Molecular Phylogeny: Influence of Nonindependence Among Sites

NICOLAS GALTIER
CNRS UMR 5000, Genome, Populations, Interactions, Université Montpellier 2, CC 63, Place E. Bataillon, 34095 Montpellier, France;
E-mail: galtier@univ-montp2.fr

Abstract.—The influence of nonindependence among sites on phylogenetic reconstructions and bootstrap scores was investigated both analytically and empirically. First, the sampling properties of the bootstrap support in the four-species case was derived for the maximum-parsimony method, assuming either independently or nonindependently evolving sites. The influence of various models of departure from the independence assumption was quantified. Second, trees and bootstrap scores estimated from subsets of consecutive (potentially coevolving) versus dispersed (presumably independent) sites of a ribosomal RNA data set were contrasted. The two approaches consistently suggest that a departure from the assumption of independent sites tends to reduce the amount of phylogenetic information contained in the data, but to increase the apparent statistical support for reconstructed trees, as measured by the bootstrap. In particular, nonindependence can lead to strongly supported wrong internal branches. [Bootstrap; coevolution; molecular phylogeny; ribosomal RNA; site independence.]

Independence among sites is a fundamental assumption of virtually all methods used to access molecular phylogeny, including model-based methods for reconstructing trees and methods for assessing the level of confidence in reconstructions, among which the bootstrap (Felsenstein, 1985) is by far the most widely used. However, biologists agree that nucleic acid and protein sites probably do not all evolve independently, mostly for functional reasons. The set of allowed (or probable) amino acid or nucleotide states at a certain site, for instance, might be constrained by the state at some other sites, e.g., neighboring sites. This is especially true for structural RNAs (e.g., ribosomal and transfer RNAs) whose secondary structure is stabilized by Watson–Crick interactions between pairs of ribonucleotides. Such paired nucleotides evidently coevolve through compensatory substitutions; only C,G; A,U, and possibly G,U pairs are allowed (Vawter and Brown, 1993; see Hickson et al., 1996, for exceptions). Specific Markov models of sequence evolution have been built to represent this peculiarity of ribosomal RNA (rRNA) data (Muse, 1995; Zhetsky, 1995; Tillier and Collins, 1995, 1998). Compensatory substitutions also have been detected in protein-coding genes (Zhang and Rosenberg, 2002).

How much a departure from the assumption of independent evolution among sites will affect phylogenetic reconstructions (when it is neglected) is empirically addressed by Tillier and Collins (1995) when they simulated the evolution of nucleotide sequences under nonindependence and assessed the performance of the neighbor-joining (NJ; Saitou and Nei, 1987) and maximum-likelihood (ML; Felsenstein, 1981) methods. ML was more robust to nonindependence than was NJ. The performance of NJ was improved when a model accounting for nonindependent changes was used. The authors commented that a departure from the assumption of independence among sites is essentially equivalent to a reduction of the number of sites (Tillier and Collins, 1995, 1998), i.e., the phylogenetic signal is reduced because several sites carry the same (or correlated) information.

Another and perhaps more worrying consequence of nonindependent sites concerns the evaluation of the reliability of the reconstructed trees. Tillier and Collins (1995) noted that incorporation of nonindependence leads to an overestimation of the statistical confidence in the trees, as measured by Tajima’s (1992) distance-based test. They argued that a similar overestimation of the level of confidence should apply for other measures, including the bootstrap, an argument previously made by Dixon and Hillis (1993). Nonindependence among sites, although decreasing the phylogenetic content of the data set, tends to increase its self-consistence, i.e., coevolving sites tend to “agree.” Consider the extreme case of a 100-site data set that includes a subset of 90 fully linked sites (i.e., sites undergoing only simultaneous changes). These 90 sites are just as reliable as any of the remaining 10 sites, but an internal branch supported by the 90 linked sites would be given a high bootstrap support value (typically >90%), whether the branch is correct or not.

This upward bias of the bootstrap support value, although strongly suggested by this example, has never been quantified theoretically or empirically, which might be the reason this bias has hardly ever been acknowledged when discussing the statistical support of trees reconstructed from, say, ribosomal versus protein data (although the former presumably violate the independence assumption to a larger extent than do the latter).

Here, I examine the influence of departure from independence on bootstrap scores, both analytically (in the simple case of a four-species maximum parsimony reconstruction) and empirically (using rRNA data).

THEORY

Definitions

Let A, B, C, and D be four taxa, and let $T = ((A,B),(C,D))$ be the true tree connecting these taxa. Let $D$ be a four-sequence data set with $n$ sites. The unweighted maximum-parsimony (MP) method will unambiguously reconstruct $T$ if and only if the number $n_1$ of sites of
the form (A:X, B:X, C:Y, D:Y) is higher than \( n_2 \) and \( n_3 \),
the number of sites of the form (A:X, B:Y, C:X, D:Y) and
(A:X, B:Y, C:Y, D:X), respectively, where X and Y are any
distinct character states. Sites not matching these three
patterns are not informative. Their total number will be
called \( n_4 \), where the sum of \( n_i \) for \( 1 \leq i \leq 4 \) is \( n \). From the
point of view of MP, \( D \) is fully described by the vector
\( (n_1, n_2, n_3, n_4) \).

Consider the process of sequence evolution that gives
rise to a data set \( D \). Usually, such a process is modeled
by defining branch lengths and a Markov transition ma-
trix. The probability of occurrence of a certain data set \( D \)
is a function of these parameters (Felsenstein, 1981). In
the four-species MP case, what matters is just the relative
proportions of the four categories of sites defined above.
Let \( p_1, p_2, p_3, \) and \( p_4 \) be the expected proportions of the
categories of sites. The \( p_i \) values sum to 1. They are
(not given here) functions of the branch lengths and of
the rate matrix. \( D = (n_1, n_2, n_3, n_4) \) follows a multino-

\[ B = \frac{1}{m} \sum_{i=1}^{m} X^i. \]

**Sampling Properties of Bootstrap Support**

(Independent Sites)

In this section, I investigate the sampling prop-er-
ties of \( B \) as a function of \( n \) and the \( p_i \) values un-der
the assumption of independent sites. First consider
the distribution of \( B \) conditional on certain data set
\( D = (n_i) \):

\[ B \mid D = B \mid (n_i) \] is just a binomial \( B(m, \Pr(X \mid (n_i))] \). The
nonconditional expectation is obtained by summing over
every possible data sets \( D = (n_i) \):

\[ E(B) = \Pr(T = T) = \sum_{n_1 \geq n_2 \geq n_3} \frac{n!}{n_1! n_2! n_3!} \prod_{i=1}^{4} p_i^{n_i}, \]

where \( \hat{T} \) is the MP tree.

The next step is to bootstrap data set \( D \). Bootstrapping
\( D \) means generating \( m \) pseudo-data sets, each one
being obtained by resampling \( n \) sites with replace-
ments out of the \( n \) sites in \( D \). Let \( D_i^* = (b_1^*, b_2^*, b_3^*, b_4^*) \)
be the \( i \)th pseudo-data set, where \( b_j^* \) is the number
of sites of category \( j \) in pseudo-data set \( i \). Let \( X^i \) be the
boolean variable defined by \( X^i = 1 \) if \( b_1^* > b_2^* \), \( b_3^* > b_4^* \),
\( X^i = 0 \) otherwise. Given \( D, X^i ' s \) are independent and
dividually distributed Bernoulli variables. The boot-
strap support for true tree \( T \) is a random variable
defined as

\[ B = \frac{1}{m} \sum_{i=1}^{m} X^i. \]

**Monte Carlo Approximations**

Equations 5 and 7 give the expectation and cumulative
distribution of the bootstrap support for the true tree in
the four-species MP case under the assumption of inde-
pendence among sites. Unfortunately, the complexity of
the calculations (\( O(n^6) \) for Eq. 5) makes them intractable
for large \( n \); thus, a Monte Carlo approximation is used.
Equation 5 can be approximated by

\[ E(B) = \Pr(B > x) = \sum_{n_1 \geq n_2 \geq n_3} \frac{n!}{n_1! n_2! n_3!} \prod_{i=1}^{4} \Pr([B \mid (n_i)] > x), \]

where \( B \mid (n_i) \) is the bootstrap support conditional on
data set \( D = (n_i) \). Because \( B \mid (n_i) \) is binomial, numer-
cal algorithms for its cumulative distribution probability
are available, allowing calculation of Equation 7. Similar
equations have been derived by Zharkikh and Li (1992a,
1992b) using a different formalism.
level of approximation can be reached:

\[ E(B) = E(X) \approx \frac{1}{a} \sum_{k=1}^{a} \left[ \frac{1}{b} \sum_{l=1}^{b} I[b_1(k, l) > b_2(k, l), b_1(k, l) > b_3(k, l)] \right], \]

where \( I = 1 \) if the indicated condition is true and zero otherwise and \( B(k, l) = [b_1(k, l), b_2(k, l), b_3(k, l)] \) is a compound multinomial deviate, i.e., a multinomial deviate \( M(n, n_i(k)/n) \), where \( n_i(k) \) \((1 \leq i \leq 4)\) is a multinomial deviate \( M(n, p_i) \). Similarly, Equation 7 will be approximated by

\[ \Pr(B > x) \approx \frac{1}{a} \sum_{k=1}^{a} \Pr([B \mid n_i(k)] > x), \]

where \( N(k) = [n_i(k)] \) \((1 \leq i \leq 4)\) is a random multinomial deviate \( M(n, p_i) \). In this study, \( a \) and \( b \) were fixed to 5,000.

**Bootstrap Support for the MP Tree**

The above equations approach the distribution of the bootstrap support for true tree \( T \), usually unknown. The distribution of \( \hat{B} \), the bootstrap support for MP tree \( \hat{T} \), might be more relevant in practice and is obtained by slight modifications of Equations 5 and 9. The condition of success for \( X \) moves from \((b_1 > b_2 \text{ and } b_1 > b_3)\) to \( \{i_{\max}(b_1, b_2, b_3) = i_{\max}(n_1, n_2, n_3)\} \), where \( i_{\max} \) returns the rank \((1, 2, \text{or } 3)\) of the highest number within parentheses. The expectation becomes

\[ \hat{E}(\hat{B}) = \sum_{n_1, n_2, n_3} \frac{n!}{n_1! n_2! n_3!} \sum_{i_{\max}(b_1, b_2, b_3) = i_{\max}(n_1, n_2, n_3)} \frac{n!}{n_1! n_2! n_3!} (n_i/n)^{b_i}, \]

which can be approximated by

\[ \hat{E}(\hat{B}) \approx \frac{1}{a} \sum_{k=1}^{a} \left[ \frac{1}{b} \sum_{l=1}^{b} I[i_{\max}(b_1(k, l)] = i_{\max}[n_i(k)] \right], \]

(and similarly for the cumulative distribution probability).

**Nonindependence Among Sites**

I now introduce departures from the assumption of independence among sites. First consider an extreme case in which the data set \( D \) includes \( n/2 \) independent pairs of fully linked sites (i.e., the two sites of any given pair must belong to the same category). Such a data set includes the same amount of information as a data set with \( n/2 \) independent sites. The probability of correct unambiguous MP reconstruction is obtained by replacing \( n \) with \( n/2 \) in Equation 1. Similarly, the expected bootstrap support (Eq. 5) becomes

\[ \hat{E}(\hat{B}) = \sum_{n_1, n_2, n_3} \frac{(n/2)!}{(n/2)!} \prod_{i=1}^{4} \frac{n_i}{n} b_i \prod_{i=1}^{4} (n_i/n)^{b_i}, \]

where \( n_1, n_2, \) and \( n_3 \) have to be even. Equation 13 can be approximated by Monte Carlo simulations as in Equation 5. Equation 9 applies, where \( B(k, l) = [b_1(k, l), b_2(k, l), b_3(k, l)] \) \((1 \leq i \leq 4)\) is a random multinomial deviate \( M(n, 2n_i(k)/n) \) and where \( n_i(k) \) \((1 \leq i \leq 4)\) is a random multinomial deviate \( M(n/2, p_i) \). The cumulative distribution probability of \( B \) under the hypothesis of \( n/2 \) pairs of fully linked sites can be obtained similarly. The modifications introduced in Equations 11 and 12 still apply here, allowing calculation of the distribution of \( \hat{B} \), the bootstrap support for the MP tree, under independence.

Other departures from independence among sites can be accommodated using equations similar to Equation 13 or their Monte Carlo approximations, as soon as a procedure for randomly generating \( D = (n_i) \) under the desired model is available. Models investigated here include (1) \( n/2 \) pairs of fully linked sites, (2) \( q \) pairs of fully linked sites plus \( n - 2q \) independent sites \((q < n/2)\), and (3) a set of \( r \) fully linked sites plus \( n - r \) independent sites \((r < n)\). A program implementing the above formulae is available on request.

**RESULTS**

The above equations make it possible to evaluate the statistical properties of the four-species MP method and of the bootstrap support under any model, where a model is determined by \( p_i \), the proportion of “good” sites, \( p_2 \) and \( p_3 \), the proportions of misleading sites, and \( n \), the total number of sites.

**Bootstrap Support and Probability of Success**

The relationship between bootstrap support and MP probability of success was first investigated using Equations 1 and 11. The proportion \( p_1 \) of “good” sites was varied between 0.05 and 0.12, and the proportions \( p_2 \) and \( p_3 \) of misleading sites were fixed to 0.05 (500 independent sites and 1,000 bootstrap replicates were used). When \( p_1 = 0.05 \), the three possible topologies are (on average) equally supported by the data, and the support for true tree \( T \) becomes high for high values of \( p_1 \). The extreme situations simulated here correspond to data sets expected under the Jukes–Cantor (1969) model of evolution along (1) a starlike tree, with an internal branch length of zero and terminal branch lengths of 0.21 substitutions/site \((p_1 = p_2 = p_3 = 0.05)\), and (2) a symmetric tree with an internal branch length of 0.16 and terminal branch lengths of 0.2 \((p_1 = 0.12, p_2 = p_3 = 0.05)\).

The expected bootstrap support for true tree \( T \) was plotted against the MP probability of success under the
hypothesis of independent sites (Fig. 1a, open squares). The probability of success varies from 0.30 to nearly 1 when \( p_1 \) varies from 0.05 to 0.012, as expected. The bootstrap support appears a good estimator of the probability of success for low and high values. For intermediate probabilities of success, the bootstrap support gives an underestimate of the actual value. If, for example, \( p_1 = 0.075 \), then 89% of the data sets would unambiguously support the true tree \( T \), but the average bootstrap support for \( T \) (averaged over every possible data set) would be 77.5%.

When the expected bootstrap support for MP tree \( \hat{T} \), not true tree \( T \), is considered, the correlation with MP probability of success is even worse (Fig. 1b, open squares). \( E(\hat{B}) \) grossly overestimates the probability of success for low \( p_1 \) but underestimates it for larger \( p_1 \). When \( p_1 = 0.05 \) (i.e., equal average support for every three alternative topologies), for example, the most parsimonious tree reconstructed from a data set generated under \( p_1 = p_2 = p_3 \) would get an average 61.5% bootstrap support. When \( p_1 \) is large, the MP tree is essentially always equal to the true tree \( T \), and the plot becomes identical to that of Figure 1a, where \( E(\hat{B}) \) underestimates the MP probability of success. These results are largely compatible with those reported by Zharkikh and Li (1992a, 1992b) and Hillis and Bull (1993) and discussed by Felsenstein and Kishino (1993), Berry and Gascuel (1996), and Efron et al. (1996).

### Influence of Nonindependence Among Sites

Nonindependence was incorporated in the above analysis by switching from a data set of 500 independent sites to a data set of 1,000 sites, composed of 500 pairs of fully linked sites (where the two sites of a given pair must belong to the same category, either correct or misleading). Data sets generated this way contain the same amount of information as data sets of 500 independent sites and yielded an equal MP probability of success for a given \( p_1 \) value. The bootstrap support (both of the true tree and the MP tree) was increased when redundancy was incorporated, as anticipated by Tillier and Collins (1995), but the increase was low: typically 0–5% for \( E(\hat{B}) \) and 0–10% for \( E(B) \) (Fig. 1, solid squares).

Additional calculations were conducted to assess the potential influence of this upward bias on phylogenetic studies. The total number of sites was fixed to 500, \( p_1 \) was set to 0.075, and \( p_2 = p_3 \) were set to 0.05. The level of departure from independence was varied by assuming that a certain fraction of the sites are fully linked pairs of sites. When this fraction was zero, the data sets met the independence assumption (500 independent sites). When the fraction was 1, the data set included 250 pairs of fully linked sites. Intermediate levels of nonindependence were also investigated, e.g., 100 pairs of fully linked sites plus 300 independent sites (when the proportion of paired sites is 0.4). Figure 2a displays the behavior of three variables as a function of the proportion of paired sites. As expected, the MP probability of success (open diamonds) decreases when the proportion of paired sites increases because the phylogenetic signal is reduced when sites are duplicated. The average bootstrap support for the MP tree, however, remains stable (solid circles). Although they are less informative, nonindependent data sets appear as self-congruent as independent data sets. The bootstrap support does not reflect the loss of phylogenetic information induced by nonindependence among sites. The third variable examined is the probability of MP success given that the bootstrap support is higher than some arbitrary threshold, namely 80%. Here, the question is how correct are the trees firmly supported by the bootstrap? This probability is close to 1 (0.99) for independent sites but drops to 0.92 when nonindependence is assumed: for data sets including 250 pairs of fully linked sites, an MP tree supported by >80% bootstrap will be wrong with a probability of 0.08. Nonindependence increases the chances of obtaining high bootstrap supports for wrong trees.

Another model of departure from the independence assumption was investigated (Fig. 2b). Here, a 250-site data set included a single group of fully linked sites. The size of this group was varied from zero (independent sites) to 25 (in which case each site of the group depends on the 24 others). The remaining sites were independent of each other. The influence of this kind of departure
Figure 2. Bootstrap support versus level of nonindependence (strong phylogenetic signal). (a) 500 sites. The x-axis gives the proportion of pairs of fully linked sites. (b) 250 sites. The x-axis gives the size of one single group of fully linked sites. Open diamonds = MP probability of success; solid circles = expected bootstrap support for the MP tree; solid triangles = MP probability of success given that the bootstrap score is >80%.

from independence appears higher than in the previous case. The bootstrap support increased when nonindependence was introduced, although the phylogenetic signal (MP probability of success) decreased. The probability that a well-supported MP tree is wrong increased from 4% (independent sites) to 24% (25 fully linked sites).

Similar calculations were performed assuming equal probabilities for the three categories of informative sites ($p_1 = p_2 = p_3 = 0.075$), i.e., no phylogenetic signal (Fig. 3). The bootstrap support of the MP tree increased with nonindependence, although the probability of success of MP was, as expected, invariably equal to 1/3. The probability that the MP tree is supported by a >80% bootstrap value was approximately 0.15 in case of independent sites but reached 0.34 for data sets including highly correlated sites (Fig. 3, open squares). The apparent phylogenetic signal, as measured from bootstrap scores, was artificially increased by nonindependence.

Data Analysis

The above analysis is restricted to the MP method and to four-species data sets. An empirical analysis of rRNA data was conducted to examine the influence of nonindependence on bootstrap support in a more general case. The rationale of this study was to contrast phylogenetic reconstructions and bootstrap scores obtained from data subsets of consecutive (potentially departing the independence assumption) versus dispersed (presumably independent) sites of a large data set.

The aligned large subunit rRNA sequences of 40 eukaryote species were obtained from the European rRNA database (http://oberon.rug.ac.be:8080/rRNA). Species were chosen according to two criteria: sequences had to be as complete as possible, and the sampling of eukaryotic phyla had to be balanced (Fig. 4). Alignments were slightly modified by eye and then searched for conserved sequence tracts of 100 nucleotides. Ten such segments of 100 unambiguously aligned sites were found (after sites including gaps or undetermined nucleotides were removed). The remaining sites were discarded. A data set of 1,000 sites remained and was designated the full data set (available on request). An MP tree was constructed from this data set was designated the full data tree.

Site subsets of the full data set were then considered. The first subsets were the 10 stretches of 100 consecutive sites, designated the WINDOW data subsets. The second group of subsets was obtained by sampling nonconsecutive sites from the full data set. Ten such data
FIGURE 4. NJ analysis (Kimura distances) of 40 eukaryote large subunit rRNA sequences with 1,000 unambiguously aligned sites. Scale bar is in unit of per site average substitution rate.

subsets were constructed, the $i$th of which included sites of ranks $i, i + 10, i + 20, \ldots, i + 990$ in the full data set; these data subsets were designated MODULO. Therefore, the 10 WINDOW and 10 MODULO data subsets each consisted of 100 sites. For each WINDOW and MODULO data subset, an MP tree was built and the bootstrap supports of reconstructed internal branches were assessed from 500 replicates. The pooled WINDOW
results were compared with the pooled MODULO results.

Under the assumption of independent sites, the location of sites in the molecule should not matter, and the two categories of data subsets should behave similarly. However, this was not the case. The accuracy of tree reconstructions was first measured by comparing the trees obtained from data subsets with the full data tree. This comparison is not an absolute measure of accuracy because the full data tree is probably not the same as the true tree. However, this comparison gives information about the potential spatial structure of the phylogenetic information of the data set. The average proportion of “correct” internal branches (i.e., internal branches shared by the full data tree) was higher for the MODULO (0.292) than for the WINDOW (0.189) data subsets (NJ analysis). The phylogenetic signal appears lower in subsets of consecutive sites than in subsets of dispersed sites.

Is this difference reflected by lower average bootstrap support for WINDOW than for MODULO data sets? Results are summarized in Table 1, separating the bootstrap supports of “correct” (i.e., shared by the full data tree) versus “incorrect” internal branches. Internal branches reconstructed from WINDOW data subsets have higher average bootstrap support than do those reconstructed from MODULO data subsets. Nine “incorrect” internal branches were supported by bootstrap scores >60% in WINDOW reconstructions, whereas three such strongly supported “incorrect” branches arose from MODULO data subsets. Consistent results were obtained when the NJ method (Saitou and Nei, 1987; Kimura’s, 1980 distances) rather than MP was used to build trees (Table 2), indicating that this problem is not restricted to MP. The average bootstrap support was substantially higher in the NJ than in the MP analysis; therefore, the threshold for “strongly supported” internal branches was set to 80%, not 60%, in the NJ analysis. Similar results were obtained when the NJ tree was constructed using distances calculated under the assumption of gamma-distributed rates across sites (Jin and Nei, 1990; assumed gamma shape parameter = 0.5; results not shown).

Subsets of consecutive sites appear to contain a lower phylogenetic signal than subsets of dispersed sites, yet they yield a higher average bootstrap score and a higher number of strongly supported dubious internal branches. These results are fully consistent with the predictions of the above theoretical study and suggest that neighboring rRNA sites do not evolve independently and that the departure from independence has some impact on tree reconstructions and bootstrap scores.

This conclusion was confirmed by an additional analysis involving a third kind of data subset called DOUBLED. These data subsets were constructed by introducing nonindependence in the MODULO subsets. A number (k) of sites (0 < k ≤ 50) were randomly sampled (without replacement) out of the 100 sites of each MODULO subset. Each of these k sites was then duplicated. Then, 100 – 2k sites (randomly sampled without replacement from the remaining 100 – k sites of the MODULO data set) were added, resulting in 10 DOUBLED data subsets. This procedure was followed twice, using k = 25 (third column of Tables 1, 2) and k = 50 (fourth column of Tables 1, 2), respectively. The DOUBLED data subsets were analyzed as described above. When 25 sites had been duplicated so that half of the sequence length corresponds to paired sites, doubled dispersed sites appeared to mimic the behavior of consecutive sites (Table 1, 2). The proportion of “correct” internal branches, the average bootstrap support for “correct” and “incorrect” branches, and the number of strongly supported “incorrect” branches in the DOUBLED (k = 25) data subsets were essentially similar to those obtained from the WINDOW data subsets. Doubling 50 sites (so that all sites of the data sets are paired) resulted in an excess of “incorrect” branches and an elevated average bootstrap support (Table 2) compared with the WINDOW data subsets. The data, therefore, appear compatible with roughly half of the sites being paired, which is typically the approximate proportion of stems in an rRNA molecule.

**DISCUSSION**

The statistical properties of the bootstrap support in the simple four-species MP case were derived analytically (following Zharkikh and Li, 1992a, 1992b) and extended to the nonindependent case. A Monte Carlo numerical approximation was proposed to overcome the high complexity of the actual formula, which allows
quick examination of the behavior of the bootstrap score under various conditions without simulating sequence data. Each data point of Figures 1, 2, and 3 typically required 10–15 minutes of computing time on a standard PC.

Whether these equations can be extended to more than four species or to other tree-building methods is a challenging issue. This extension appears feasible in theory but probably not in practice. In any case, the bootstrap score for certain internal branches is a function of a compound multinomial distribution. In the four-species MP case, the multinomial has four classes, and the function simply involves calculating the maximum of the number of outcomes in three of the four classes. Adding species would increase the number of relevant site categories, i.e., the number of multinomial classes. Using distance-based or likelihood methods with complex evolutionary models would also increase this number (because sites AAGG and AACC, for example, would have to be distinguished) and would involve a function much more complex than the maximum over certain categories. Given the already high complexity of the formulæ in the four-species MP case, it seems unlikely that an analytical approach could be useful in practice in more complex cases. Simulations (Hillis and Bull, 1993) or empirical analyses (this study) appear more appropriate.

When various forms of departure from independence among sites were incorporated, the probability of MP success decreased, but this drop of phylogenetic signal was not reflected by bootstrap scores; the bootstrap support of the MP tree stagnated, or even increased, under nonindependence in agreement with the prediction of Tillier and Collins (1995). Nonindependence tends to increase the probability of strongly supported wrong reconstructions and tends to generate apparent signal when there is none. Overall, departure from independence results in an overestimation of the confidence to be put in reconstructions, as measured by the bootstrap.

An analysis of rRNA data was fully consistent with these theoretical results. Subsets of consecutive sites (potentially not independent) yielded less accurate phylogenetic reconstructions (as measured from the full data set) but higher bootstrap scores than did subsets of dispersed sites. Incorporating nonindependence by duplicating a certain proportion of dispersed sites mimicked the behavior of consecutive sites.

Should we, therefore, worry about past and future interpretations of bootstrap scores? Probably not too much because the influence of nonindependence on the bootstrap score appears slight. The overestimation was always <10% in Figure 1. The number of strongly supported dubious internal branches in WINDOW data subsets also was low. Given the highly empirical use of bootstrap scores in current practice (i.e., what should be considered a significant bootstrap score?), accounting for nonindependence effects probably would not have changed much of the molecular phylogeny literature, although nonindependence occasionally might have encouraged misleading interpretations. This additional methodological bias should be kept in mind when conducting a molecular phylogenetic analysis.

However, the WINDOW versus MODULO contrast is conservative with respect to detecting coevolutionary effects. Some rRNA stems are formed by paired RNA segments more distant from each other (in the primary structure) than the arbitrary 100-nucleotide threshold used in this study. Such long-distance interacting sites fall in distinct WINDOW data sets but might co-occur in MODULO data sets, lowering the power of the analysis.

The consecutive versus dispersed sites approach appears a suitable empirical way to roughly quantify the amount of nonindependence in actual data sets. Application to various protein data sets, for example, might yield valuable insights about actual processes of sequence evolution. However, the methodology must be improved. In particular, it is unclear when the behavior of two kinds of data sets (e.g., WINDOW vs. MODULO) is significantly different between the statistics used in this study (proportion of “correct” internal branches, average bootstrap support) involve averaging over nonindependent measures; distinct branches of a reconstructed tree and their bootstrap scores are not independent from each other.

ACKNOWLEDGMENTS

This work was supported by the Genopole Montpellier Languedoc-Roussillon and the French Appel d’offre Bioinformatique Inter-EPST.

REFERENCES


First submitted 18 March 2003; reviews returned 11 June 2003; final acceptance 14 September 2003

Associate Editor: Nick Goldman