

## ACCURACY OF NEIGHBOR JOINING FOR $n$ -TAXON TREES

KORBINIAN STRIMMER AND ARNDT VON HAESLER

Zoologisches Institut, Universität München, Postfach 202136, D-80021 München, Germany;  
E-mail: arndt@zi.biologie.uni-muenchen.de (A.vH.)

*Abstract.*—A Monte Carlo approach was used to estimate the accuracy of a given tree reconstruction method for any number of taxa. In this procedure, we sampled randomly over all possible bifurcating trees assigning substitution rates (branch lengths) to each edge from an exponential distribution to obtain a biologically sensible maximal observed distance. Three different sets of trees were studied: the unrestricted tree space, the biologically meaningful tree space as introduced by Nei et al. (1995, *Science* 267:253–254), and the population data tree space. We used this technique to elucidate the performance of neighbor joining as a function of the number of taxa, assuming that distances are uncorrected and sequences evolve according to the Jukes–Cantor model. The accuracy of neighbor joining decreases almost exponentially with the number of taxa. However, the rate of decrease depends on the tree space studied. Although the accuracy decreases towards zero, the similarity, i.e., the number of partitions that are identical between model tree and reconstructed tree, is in all cases studied much higher than the value expected for two randomly chosen trees. Although the probability of recovering the true tree is dramatically influenced by sequence length, the average similarity does not decrease substantially if branch lengths are not too short. [Assigning edge lengths; Felsenstein zone; finite sequence length; Jukes–Cantor model; Monte Carlo sampling; neighbor joining.]

A great deal of work has been put into studies on the accuracy of tree reconstruction methods based on DNA or amino acid sequences, where accuracy is understood as the ability of a reconstruction method to recover the true branching pattern (topology) of the underlying tree from a given data set. If one wants to understand the accuracy of various tree building methods, in principle, one has to consider the whole space of trees, i.e., all possible topologies with all possible assignments of edge lengths. More formally, let  $T_n$  be the set of all nondegenerate binary trees with  $n$  leaves (taxa, species, sequences). The size of  $T_n$  rapidly increases with  $n$ , and the number of trees  $|T_n| = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2n - 5)$ . Each tree  $T \in T_n$  has exactly  $2n - 3$  branches, and each branch is assigned a number of substitutions between zero and infinity. We denote with  $\mathcal{T}_n$  the space of all different tree topologies plus assignments of edge lengths. If one wants to understand the behavior of a tree reconstruction method for a given evolutionary model, one therefore ought to study the entire space  $\mathcal{T}_n$  or an appropriately defined region thereof.

There have been two major approaches

giving more insight into this huge space, each with its own simplifying assumptions. The first approach confines itself to a four-taxon tree with only two different edge lengths assigned to the five branches of the tree. This enables one to examine the accuracy of different tree reconstruction methods under a variety of evolutionary models (Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995). This approach facilitates determination of the region in  $\mathcal{T}_n$  where a method fails to reconstruct the correct topology even if infinitely long sequences are used. This region, also called the Felsenstein zone, depends on the tree reconstruction algorithm and the model of sequence evolution. The main merit of this approach is the fast comparison of alternative tree reconstruction methods. The finding that distance-based reconstruction methods show a nonempty Felsenstein zone if the observed sequence differences are not properly corrected is based on this type of analysis (Hillis et al., 1994; Schöniger and von Haeseler, 1995).

The second approach is to analyze the accuracy of tree reconstruction methods for cases with more than four taxa and with more than two different edge lengths

by considering trees whose edge lengths follow a molecular clock or specific trees where this assumption is severely violated (Schöniger and von Haeseler, 1993). These studies have been useful, for example, in estimating the length of sequences necessary to obtain the true topology (Tateno et al., 1982, 1994; Tateno and Tajima, 1986; Saitou and Nei, 1987; Sourdis and Krimbas, 1987; Sourdis and Nei, 1988; Saitou and Imanishi, 1989; Jin and Nei, 1990; Schöniger and von Haeseler, 1993; Kuhner and Felsenstein, 1994). However, one should be aware that it is not necessarily possible to apply the results obtained for the few model trees to the whole space of trees  $\mathcal{T}_n$ .

In this study, we utilized a third approach to evaluating the behavior of tree reconstruction methods for the space  $\mathcal{T}_n$ . Specifically, we examined the accuracy of the neighbor-joining method (Saitou and Nei, 1987) if sequences evolve according to the Jukes–Cantor model of sequence evolution (Jukes and Cantor, 1969) but uncorrected distances are used in the analysis. Neighbor joining can be inconsistent in this situation (Huelsenbeck and Hillis, 1993). However, although this scenario is an oversimplification of the true evolutionary history and of the way in which phylogenetic analysis is usually conducted, it still may elucidate the reliability of neighbor joining when data from the real world are used. Such studies are valuable because of the widespread use of neighbor joining, especially for large data sets (Hedges et al., 1992).

We investigated the whole tree space using a Monte Carlo procedure to sample randomly across the topologies  $\mathcal{T}_n$ . In pursuing this approach, special attention must be paid to the problem of assigning branch lengths to each edge. Although it is possible to apply a generalization of the procedure previously used in studies of the four-taxon case, this method generally implies biologically unacceptable high numbers of substitutions between any two taxa. Instead, we proposed to draw the branch length along a particular edge from an exponential distribution in such a way

that the maximal observed distance is biologically meaningful. With this procedure, it is possible to estimate a meaningful volume of the Felsenstein zone for any number of taxa.

Herein, we describe the principle of the Monte Carlo procedure employed to study the accuracy of tree reconstruction methods for any number of taxa, giving special attention to the problem of assigning edge lengths to the branches of a tree. The Monte Carlo procedure was applied to the neighbor-joining method to determine the accuracy of neighbor joining and to compute the average number of identical partitions, i.e., a measure of similarity, between original tree and inferred tree. These calculations were done for three different regions of the tree space: the unrestricted tree space, the biologically meaningful region as introduced by Nei et al. (1995), and the population data region. Simulations were done for infinitely long sequences and sequences of finite length. These results allow us to estimate the reliability of neighbor joining in recovering complex topologies from short sequences, as is common practice, for example, in studies on the origin of modern humans (Hedges et al., 1992).

## METHODS

### *Model of Sequence Evolution*

We assume that sequences evolve according to the model of Jukes–Cantor (Jukes and Cantor, 1969). If  $d$  substitutions per site occurred between two sequences, then the observed number of differences  $\tilde{d}$  per site is

$$\tilde{d} = \frac{3}{4}(1 - e^{-(4/3)d}). \quad (1)$$

We differentiate with a tilde the observed and expected value of the same entity. As  $d$  approaches infinity,  $\tilde{d}$  reaches the so-called saturation level,  $3/4$ . The nonlinear relationship between  $d$  and  $\tilde{d}$  is the reason why distance-based methods without proper correction can fail to reconstruct the correct tree topology even if sequence length is infinite.

Let us now assume that taxon A and B are linked by two branches. Let  $a$  and  $b$  be the expected number of substitutions along these two branches. We denote the probability of observing different nucleotides at the two end nodes of each branch with  $\tilde{a}$  and  $\tilde{b}$ . These probabilities are computed using Equation 1. Although the number of substitutions is perfectly additive

$$d_{AB} = a + b, \quad (2)$$

the number of observed differences  $\tilde{d}_{AB}$  between A and B is computed from  $\tilde{a}$  and  $\tilde{b}$  with the formula

$$\tilde{d}_{AB} = \tilde{a} + \tilde{b} - \frac{4}{3}\tilde{a}\tilde{b}, \quad (3)$$

where the last term can be viewed as a correction for multiple hits. This additional term ensures that  $\tilde{d}_{AB}$  never exceeds  $\frac{3}{4}$  if  $\tilde{a} \leq \frac{3}{4}$  and  $\tilde{b} \leq \frac{3}{4}$ . Repeated application of Equation 3 allows the computation of observed differences if two taxa are connected by more than two branches in a tree.

#### *Assigning Substitutions along a Branch*

To elucidate the performance of a specific tree reconstruction method for a given tree topology, edge lengths must be assigned. Usually, the observed branch lengths  $\tilde{a}$  vary from 0 to  $\frac{3}{4}$  (Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995). This means that observed branch lengths  $\tilde{a}$  are drawn uniformly from the interval  $[0, \frac{3}{4}]$ . However, and this has been overlooked in most studies so far, this assumption implies (transformation of density using Equation 1) that the number of substitutions  $a$  that actually occurred along each edge is on average over the considered tree space  $\mu = \frac{3}{4}$  and follows an exponential distribution with parameter  $\frac{1}{\mu}$ . This extremely high number of expected substitutions per edge is biologically unrealistic for most data. Furthermore, if we assume that two taxa are connected by  $k$  branches, then the average number of expected substitutions equals  $\mu k$ ; the observed number of differences quickly approaches saturation level. Thus, this procedure of assign-

ing edge lengths contains implicit assumptions that are in obvious conflict with real alignable data sets.

We call this space of trees showing an average number of expected substitutions  $\mu = \frac{3}{4}$  per edge the unrestricted tree space  $\mathcal{T}_n^{\text{unr}}$ . If we look at real data sets used in conjunction with a phylogenetic analysis, we see that the maximal observed distance  $\tilde{d}_{\text{max}}$  between any two taxa hardly ever exceeds a value of 0.57 (Nei et al., 1995). The trees obeying this criterion are inhabiting the so-called biologically meaningful tree space  $\mathcal{T}_n^{\text{bmr}}$ . The sequences used in population studies are even more sparse in substitutions; a worldwide sample of the hypervariable region I of the mitochondrial control-region in humans (Vigilant et al., 1991; Zischler et al., 1995) suggests  $\tilde{d}_{\text{max}} = 0.05$ . We call the corresponding tree space  $\mathcal{T}_n^{\text{pop}}$ .

The assignment of expected substitutions per edge drawn from an exponential distribution implies that in principle every branch can take on every possible expected branch length, although larger lengths are highly unlikely and the average expected number of substitutions ( $\mu$ ) remains finite. To obtain more biologically reasonable values for observed pairwise distances and to ensure at the same time that every branch can still in principle take on any expected branch length, we abandon the fixation to the unrealistically high value of  $\mu = \frac{3}{4}$  and vary  $\mu$  depending on the number of taxa to obtain the desired value of  $\tilde{d}_{\text{max}}$ . Given the number  $n$  of species in the tree, we infer the average maximal number of branches  $B_{\text{max}}$  connecting two taxa. Because there is no analytical formula for this relationship, we have generated 10,000 random bifurcating trees and averaged over the maximal number of branches of each tree. Table 1 lists some  $n$  and corresponding  $B_{\text{max}}$  values. Once  $B_{\text{max}}$  is obtained, we can easily calculate the average expected branch length  $\mu$  using the Jukes–Cantor equation

$$\mu = -\frac{3}{4B_{\text{max}}}\ln\left(1 - \frac{4}{3}\tilde{d}_{\text{max}}\right). \quad (4)$$

TABLE 1. Average expected number  $\mu$  of substitutions per edge in dependency of the number of taxa  $n$  and maximal observed pairwise distance  $\tilde{d}_{\max}$ . The average maximal number of branches  $B_{\max}$  connecting two taxa in a tree is used to compute  $\mu$ .

$n$	$B_{\max}$	$\mu$	
		$\tilde{d}_{\max} = 0.57$	$\tilde{d}_{\max} = 0.05$
5	4.00	0.2676	0.0129
10	7.86	0.1362	0.0066
20	13.55	0.0790	0.0038
30	17.97	0.0596	0.0029
40	21.79	0.0491	0.0024

The value of  $\mu$  depends on the observed maximal distance  $\tilde{d}_{\max}$  and the number of taxa  $n$ . If  $n$  increases and  $\tilde{d}_{\max}$  is held constant, the average expected number of substitutions per edge  $\mu$  has to decrease. In Table 1, we show values for  $\mu$  for some configurations of  $n$  and  $\tilde{d}_{\max}$ . In all cases,  $\mu$  is much smaller than the value of  $\frac{3}{4}$  implicitly used in the standard procedure.

The uniform distribution of observed branch lengths  $\tilde{a}$  implies an exponential distribution of expected branch lengths  $a$  with mean  $\mu = \frac{3}{4}$ . If we allow the expectation  $\mu$  to vary, we can ask what is now the corresponding distribution of observed substitutions. A simple back transformation reveals that  $\tilde{a}$  follows

$$w(\tilde{a}) = \frac{1}{\mu} \left(1 - \frac{4}{3} \tilde{a}\right)^{\frac{3}{(4\mu)} - 1}. \quad (5)$$

It is easy to show that

$$\int_0^{3/4} w(\tilde{a}) d\tilde{a} = 1. \quad (6)$$

For  $\mu = \frac{3}{4}$ , we recover the uniform distribution. Using this distribution governing  $\tilde{a}$ , we can utilize Equation 3 to compute observed distances along the tree.

Many other ways of assigning branch lengths are conceivable, e.g., drawing  $\tilde{a}$  uniformly from a restricted interval  $[0, \frac{3}{4}]$  with  $r < \frac{3}{4}$ . But this would mean that the expected branch lengths were also restricted. Our choice has the virtue of containing the standard procedure and still allowing every branch the possibility of taking every expected branch length.

### Monte Carlo Sampling across Tree Topologies

To explore the tree space and to characterize the accuracy of tree reconstruction methods, we employed the following simple scheme:

1. Generate a random tree topology, i.e., choose randomly one of the  $|T_n| = 1 \cdot 3 \cdot \dots \cdot (2n - 5)$  labeled, binary, and unrooted trees, each having probability  $1/|T_n|$ .
2. Assign length  $\tilde{a}$  to each of the  $2n - 3$  branches of the tree, where  $\tilde{a}$  is drawn from the distribution described in Equation 5 and  $\mu$  is determined from Equation 4.
3. Calculate the observed number of differences per site between all  $n(n - 1)/2$  pairs of sequences using Equation 3 repeatedly.
4. Apply the tree reconstruction algorithm (here neighbor joining) to reconstruct a tree.
5. Compare the resulting tree topology with the original tree.
6. Go back to step 1 until a fixed number of repeats are completed. In this study, we did 10,000 iterations per simulation.

In this way we are able to sample randomly across trees that belong to a given tree space defined by  $n$  and  $\tilde{d}_{\max}$ .

### Statistics

In the comparison step of the above Monte Carlo algorithm we computed the following statistics:

1. The accuracy  $\epsilon$ , i.e., the proportion of correctly reconstructed trees among the 10,000 random trees.
2. The proportion of partitions  $\beta$  that are identical in the reconstructed and original tree. A partition is defined as a split of the set of  $n$  taxa in two disjoint, non-empty subsets, induced by any of the  $n - 3$  internal edges. Only inner edges were considered, because outer edges induce trivial partitions of one taxon versus the rest. These partitions are identical in all trees. If two trees have the same topology, then the  $n - 3$  non-trivial partitions are identical and vice versa.  $\beta$  gives an estimate of the average

similarity between model trees and the reconstructed trees.

Depending on the tree reconstruction method, many other statistics are conceivable. For a clustering method like neighbor joining for example, one could estimate the probabilities of each pairing step to be the first erroneous step that destroys the identity of the tree and the reconstructed tree.

#### *Finite Sequence Length*

Another important issue is that of finite sequence length. The simple formula for Jukes–Cantor evolution (Eq. 1), describing the dependency between the expected number of differences  $\tilde{d}$  and the true distance  $d$ , is valid only for infinitely long sequences, i.e., if distances can be estimated with infinitely high precision. However, if finite sequence lengths are considered, one has to take two effects into account, the sampling error and the reduced resolution of distances. Both can be easily incorporated into the Monte Carlo procedure as we have described here. Let us consider a specific branch and assume that we have already fixed the observed branch length  $\tilde{a}$  that we have drawn from the distribution Equation 5. This  $\tilde{a}$  can also be interpreted as the probability of observing a difference for a specific site along the sequence. For a sequence with length  $l$  the actual number of differences that are observed along the sequence follows a binomial distribution with parameters  $\tilde{a}$  and  $l$ . Drawing from this distribution and dividing the actual number of observed differences by the sequence length gives us a new estimate for  $\tilde{a}$  that now also incorporates random sampling error. Finally, when the observed distances between all pairs of taxa have been calculated, we must consider the reduced resolution induced by finite sequence length. The observed distances between any two taxa can take on only the discrete values  $0, \frac{1}{l}, \frac{2}{l}, \dots, 1$ . Thus, the final distance matrix must be rounded to the grid defined by this resolution. Adopting these two methods in the simulation procedure therefore enables us to study finite sequence lengths as well.

TABLE 2. Estimated accuracies  $\epsilon$  for the unrestricted region (unr), the biological meaningful region (bmr), and the population data region (pop) for 4–15 taxa.

No. taxa	$\epsilon$ (%)		
	unr	bmr	pop
4	80.8	86.4	98.9
5	56.4	72.4	97.5
6	36.1	58.9	96.7
7	20.3	48.8	95.5
8	12.4	38.6	95.0
9	6.3	33.8	93.5
10	3.4	27.5	92.7
11	1.7	23.0	91.6
12	0.8	19.1	90.8
13	0.4	15.8	90.3
14	0.3	13.1	89.4
15	0.1	10.6	88.6

#### RESULTS: PERFORMANCE OF NEIGHBOR JOINING

We used the Monte Carlo procedures to evaluate neighbor joining (Saitou and Nei, 1987), a popular tree reconstruction method widely used in systematic biology. We investigated its performance on the three different tree spaces  $\mathcal{T}_n^{\text{unr}}$ ,  $\mathcal{T}_n^{\text{bmr}}$ , and  $\mathcal{T}_n^{\text{pop}}$  for  $n = 4$  to  $n = 150$  taxa. The computer program used (ANSI C) is available upon request from the authors.

#### *Infinite Data*

For infinitely long sequences, the accuracy  $\epsilon$  of neighbor joining, i.e., the probability of correctly reconstructing the complete tree, for up to 15 taxa is shown in Table 2. As  $n$  increases, the accuracy decreases rapidly. For a tree with only 15 taxa, the probability that neighbor joining will recover the true tree topology is practically zero if the unrestricted tree region is considered. The probability of obtaining the true tree also drops for the remaining two regions. However, the decrease is much less pronounced. The accuracy is still  $>88\%$  for 15 taxa if the population data region is analyzed.

Figure 1 displays the estimated logarithms ( $\log_{10}$ ) of accuracies  $\epsilon^{\text{unr}}$ ,  $\epsilon^{\text{bmr}}$ , and  $\epsilon^{\text{pop}}$  as a function of  $n$  ( $n = 4$ –150) for the unrestricted region, the biologically meaningful region, and the population data re-

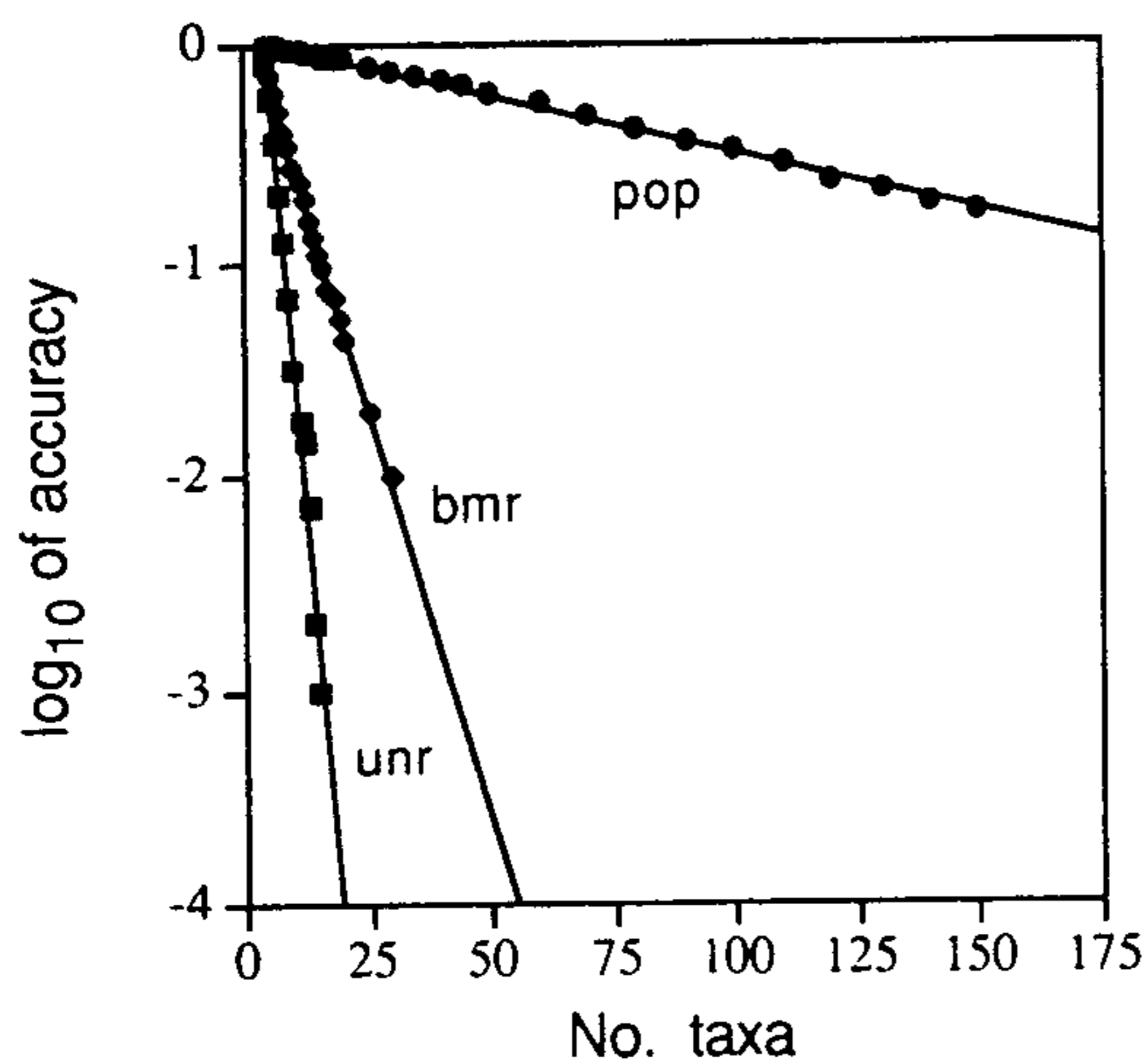


FIGURE 1.  $\log_{10}$  accuracy as function of the number of taxa for the three investigated regions of the tree space: unr = unrestricted; bmr = biologically meaningful; pop = population data.

gion, respectively. All three curves show a log-linear relation with different slopes. Although the probability of obtaining the correct tree is  $<10\%$  for  $n > 8$  or  $n > 15$  taxa if the unrestricted or biologically meaningful region, respectively, is analyzed, the accuracy of neighbor joining in the population data region is still  $17.0\%$  for 150 taxa. If one accepts something like a  $90\%$  chance of recovering the true topology as a threshold for a perfect tree building method, then neighbor joining is only applicable to  $\mathcal{T}_n^{\text{pop}}$  with at most 13 taxa.

Although the true tree is rarely retrieved, the reconstructed tree topology may be close to the true branching pattern with only a few taxa misplaced. To evaluate this possibility, we measured the similarity between the reconstructed tree and the true tree by counting the number of partitions  $S_n$  that are identical in both trees (Hendy et al., 1988). Figure 2 shows the average relative number of identical partitions  $\beta$  per tree as estimated from 10,000 of the Monte Carlo runs. Depending on the region of the space, the curves display a distinct shape.

For the biologically meaningful region, the average similarity between the reconstructed trees and the true trees is monotonically decreasing, having an average

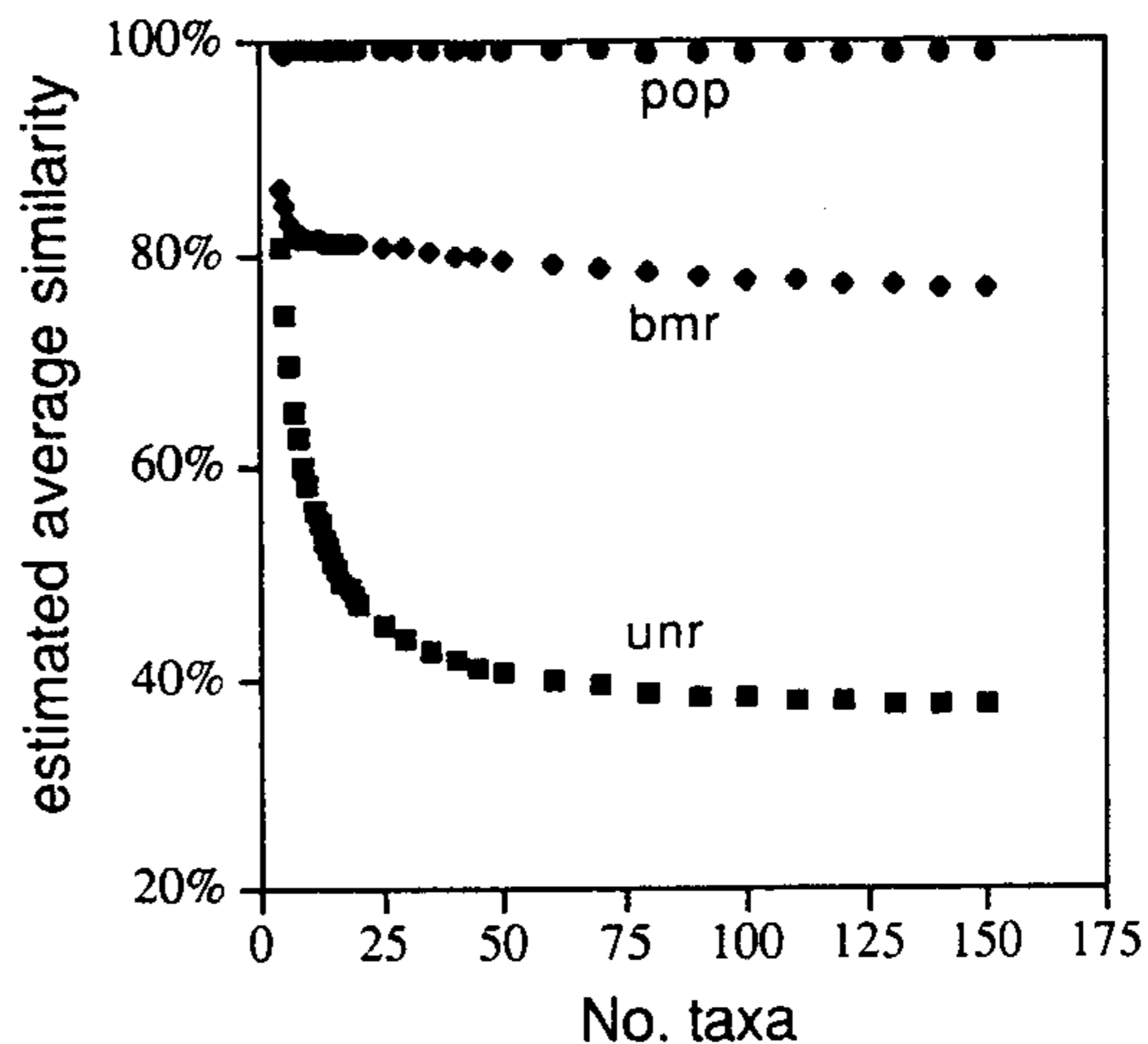


FIGURE 2. Proportion  $\beta$  of correctly reconstructed partitions on  $\mathcal{T}_n^{\text{unr}}$ ,  $\mathcal{T}_n^{\text{bmr}}$ , and  $\mathcal{T}_n^{\text{pop}}$ : unr = unrestricted region; bmr = biologically meaningful region; pop = population data region.

similarity of  $76.5\%$  for 150 taxa. For  $n = 30$  taxa, this value is  $80.6\%$ . If  $\mathcal{T}_n^{\text{unr}}$  is analyzed, the average similarity changes drastically with  $n$ . The  $\beta$  value quickly drops to around  $40\%$  ( $n = 50$ ), and then the slope of the curve is very small. For 1,000 taxa, the average similarity is only about  $35.6\%$ ; however, this value is still  $>12.5\%$ , the expected number of identical partitions for two randomly chosen trees with many taxa (Hendy et al., 1988; Steel and Penny, 1993).

#### Finite Data

We have studied the accuracy  $\epsilon$  and average similarity  $\beta$  for  $\mathcal{T}_n^{\text{bmr}}$  and  $\mathcal{T}_n^{\text{pop}}$  for various numbers of taxa  $n$  and different finite sequence lengths  $l$ . The analysis of the population data region  $\mathcal{T}_n^{\text{pop}}$  is motivated by studies about the origins of modern humans, where a tree for many taxa is reconstructed based on short sequences (Vigilant et al., 1991; Hedges et al., 1992). If  $l \approx 300$ , then the hypervariable region I of the mitochondrial D-loop may serve as an example (Zischler et al., 1995).

Table 3 summarizes our results for  $n = 30$ ,  $n = 90$ , and  $n = 150$  and for  $l = 100$ ,  $l = 300$ ,  $l = 1,000$ , and  $l = 10,000$ . Even for relatively long sequences ( $l = 1,000$ ) and only a few taxa ( $n = 30$ ), the probability of

TABLE 3. Accuracies  $\epsilon^{\text{pop}}$  and  $\epsilon^{\text{bmr}}$  (%) and percentage of correctly identified partitions  $\beta^{\text{pop}}$  and  $\beta^{\text{bmr}}$  for neighbor joining if  $T_n^{\text{pop}}$  and  $T_n^{\text{bmr}}$  are analyzed for sequences of finite length  $l$ .

$l$	No. taxa									
	30				90 <sup>a</sup>			150 <sup>a</sup>		
	$\epsilon^{\text{pop}}$	$\epsilon^{\text{bmr}}$	$\beta^{\text{pop}}$	$\beta^{\text{bmr}}$	$\epsilon^{\text{pop}}$	$\beta^{\text{pop}}$	$\beta^{\text{bmr}}$	$\epsilon^{\text{pop}}$	$\beta^{\text{pop}}$	$\beta^{\text{bmr}}$
100	0.00	0.07	26.9	74.4	0.00	14.7	67.0	0.00	11.2	62.2
300	0.00	0.40	54.3	78.5	0.00	35.8	74.3	0.00	28.6	71.7
1,000	0.61	0.58	80.6	79.9	0.00	66.6	77.0	0.00	59.1	75.2
10,000	48.51	0.72	97.3	80.4	1.69	95.3	77.8	0.00	93.9	76.4
$\infty$	74.79	0.97	98.9	80.6	35.38	98.8	77.9	16.95	98.7	76.5

<sup>a</sup> Accuracies are already zero for 90 and 150 taxa if neighbor joining is applied to  $T_n^{\text{bmr}}$  with infinitely long sequences, therefore the corresponding columns  $\epsilon^{\text{bmr}}$  are omitted.

precisely reconstructing the model tree is practically zero, irrespective of the tree space. However, whereas the accuracy is already very low for the biologically meaningful region when infinitely long sequences are used, there is a sharp decrease in accuracy for the population data space as shorter sequences are considered. The decrease of the average similarity  $\beta$  is less pronounced. If substitution rates along the edges are high, i.e., if the maximal observed distance is high or if only a few taxa are considered, the similarity remains more or less the same unless sequences are extremely short ( $l = 100$ ).

Although sequence length is important for the accuracy or the similarity if the population data space is analyzed, it is of minor importance for the biologically meaningful space. For the biologically meaningful space, the systematic underestimation of distances due to multiple substitutions governs the performance of neighbor joining, whereas for the population data space the lack of resolution of very short edges becomes a major factor for accuracy and similarity.

#### DISCUSSION

In this simulation study, we estimated the performance of neighbor joining with regard to accuracy and average similarity in dependence of the number of taxa when distances are underestimated. As a model for this situation, we evolved sequences according to a Jukes–Cantor model and used the observed distances as input for the tree reconstruction algorithm. Although this

evolutionary set-up is surely a gross oversimplification, it nevertheless exhibits features such as the almost perfect exponential decay of accuracy with the number of taxa that may be characteristic for more complex situations and other tree reconstruction methods. Furthermore, the average similarity between the true and the reconstructed tree is remarkably high even when a large number of taxa are involved. We have also drawn attention to the problem of how to assign edge lengths to model trees, an aspect that has been overlooked in most other studies of tree reconstruction methods (Huelsenbeck and Hillis, 1993; Hillis et al., 1994; Huelsenbeck, 1995; Schöniger and von Haeseler, 1995). We demonstrated that the procedure used in previous studies of the four-taxon case is generally not suitable and may produce misleading results.

Several extensions of our work are possible. More realistic models of sequence evolution could be used, other tree building methods could be analyzed, and alternative ways of assigning branch lengths are conceivable. However, these extensions may be difficult to explore because of the enormous computation time required, e.g., for reconstruction methods that attempt to optimize an objective function such as parsimony or maximum likelihood.

#### ACKNOWLEDGMENTS

Support from the Deutsche Forschungsgemeinschaft is greatly appreciated. We also thank Nick Goldman and Michael Schöniger for helpful comments and Svante Pääbo and all the members of his

group for providing a stimulating environment. We appreciate valuable comments during the review process that helped to improve this manuscript.

#### REFERENCES

- HEDGES, S. B., S. KUMAR, K. TAMURA, AND M. STONEKING. 1992. Human origin and analysis of mitochondrial DNA. *Science* 255:737–739.
- HENDY, M., M. A. STEEL, D. PENNY, AND I. M. HENDERSON. 1988. Families of trees and consensus. Pages 355–362 in *Classification and related methods of data analysis* (H. H. Bock, ed.). Elsevier, Amsterdam.
- HILLIS, D. M., J. P. HUELSENBECK, AND C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenies. *Science* 264:671–677.
- HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44:17–48.
- HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42:247–264.
- JIN, L., AND M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* 7:82–102.
- JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 in *Mammalian protein metabolism* (H. N. Munro, ed.). Academic Press, New York.
- KUHNER, M. K., AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11:459–468.
- NEI, M., N. TAKEZAKI, AND T. SITNIKOVA. 1995. Assessing molecular phylogenies. *Science* 267:253–254.
- SAITOU, N., AND T. IMANISHI. 1989. Relative efficiencies of the Fitch–Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.* 6:514–525.
- SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.
- SCHÖNIGER, M., AND A. VON HAESLER. 1993. A simple method to improve the reliability of tree reconstructions. *Mol. Biol. Evol.* 10:471–483.
- SCHÖNIGER, M., AND A. VON HAESLER. 1995. Performance of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when sequence sites are not independent. *Syst. Biol.* 44:533–547.
- SOURDIS, J., AND C. KRIMBAS. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. *Mol. Biol. Evol.* 4:159–166.
- SOURDIS, J., AND M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Mol. Biol. Evol.* 18:298–311.
- STEEL, M., AND D. PENNY. 1993. Distributions of tree comparison metrics—Some new results. *Syst. Biol.* 42:126–141.
- TATENO, Y., M. NEI, AND F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Mol. Evol.* 18:387–404.
- TATENO, Y., AND F. TAJIMA. 1986. Statistical properties of molecular tree construction methods under the neutral mutation model. *J. Mol. Evol.* 23:354–361.
- TATENO, Y., N. TAKEZAKI, AND M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* 11:261–277.
- VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, AND A. C. WILSON. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–1507.
- ZISCHLER, H., H. GEISERT, A. VON HAESLER, AND S. PÄÄBO. 1995. A nuclear fossil of the mitochondrial D-loop and the origin of modern humans. *Nature* 378:489–492.

*Received 30 October 1995; accepted 30 May 1996*

*Associate Editor: David Baum*