J Mol Evol (2000) 51:544–553
DOI: 10.1007/s002390010118

JOURNAL OF MOLECULAR EVOLUTION

Efficiency of the Neighbor-Joining Method in Reconstructing Deep and Shallow Evolutionary Relationships in Large Phylogenies

Sudhir Kumar, Sudhindra R. Gadagkar

Department of Biology, Arizona State University, Tempe, AZ 85287-1501, USA

Received: 7 March 2000 / Accepted: 2 August 2000

Abstract. The neighbor-joining (NJ) method is widely used in reconstructing large phylogenies because of its computational speed and the high accuracy in phylogenetic inference as revealed in computer simulation studies. However, most computer simulation studies have quantified the overall performance of the NJ method in terms of the percentage of branches inferred correctly or the percentage of replications in which the correct tree is recovered. We have examined other aspects of its performance, such as the relative efficiency in correctly reconstructing shallow (close to the external branches of the tree) and deep branches in large phylogenies; the contribution of zero-length branches to topological errors in the inferred trees; and the influence of increasing the tree size (number of sequences), evolutionary rate, and sequence length on the efficiency of the NJ method. Results show that the correct reconstruction of deep branches is no more difficult than that of shallower branches. The presence of zero-length branches in realized trees contributes significantly to the overall error observed in the NJ tree, especially in large phylogenies or slowly evolving genes. Furthermore, the tree size does not influence the efficiency of NJ in reconstructing shallow and deep branches in our simulation study, in which the evolutionary process is assumed to be homogeneous in all lineages.

Key words: Phylogenetic inference — Neighbor-joining method — Large phylogenies — Zero-length branches — Accuracy — Deep versus shallow branches

Introduction

The scope of molecular phylogenetic studies for inferring short- and long-term evolutionary histories of organisms and multigene families has expanded greatly beyond molecular systematics due to an explosive growth in the number of sequences available in genetic databases (e.g., Balczarek et al. 1997; Duret et al. 1994; Higgins et al. 1996; Kumar et al. 1996; Kumar and Rzhetsky 1996; Li 1997; Nei and Kumar 2000). With this growth, data sets for molecular phylogenetics have increased in terms of the number of sequences being analyzed, and the neighbor joining (NJ) method (Saitou and Nei 1987) has become one of the most commonly used methods. It is computationally efficient, has desirable statistical properties, and is known to produce trees as accurate as, or better than, more computationally intensive and global searching methods (Charleston et al. 1993; Gascuel 1994, 1997; Kuhner and Felsenstein 1994; Nei and Kumar 2000; Nei et al. 1998; Rzhetsky and Nei 1992; Tateno et al. 1994).

Computer simulations provide a convenient way to assess the efficiency of tree-making methods (reviewed by Nei and Kumar 2000). For the NJ method, most of these computer simulation studies have evaluated its overall performance in inferring phylogenetic trees by either calculating its performance in inferring the true tree topology completely or estimating the proportion of

Correspondence to: Dr. Sudhir Kumar, Life Sciences A-371, Department of Biology, Arizona State University, Tempe, AZ 85287-1501, USA; *e-mail:* s.kumar@asu.edu

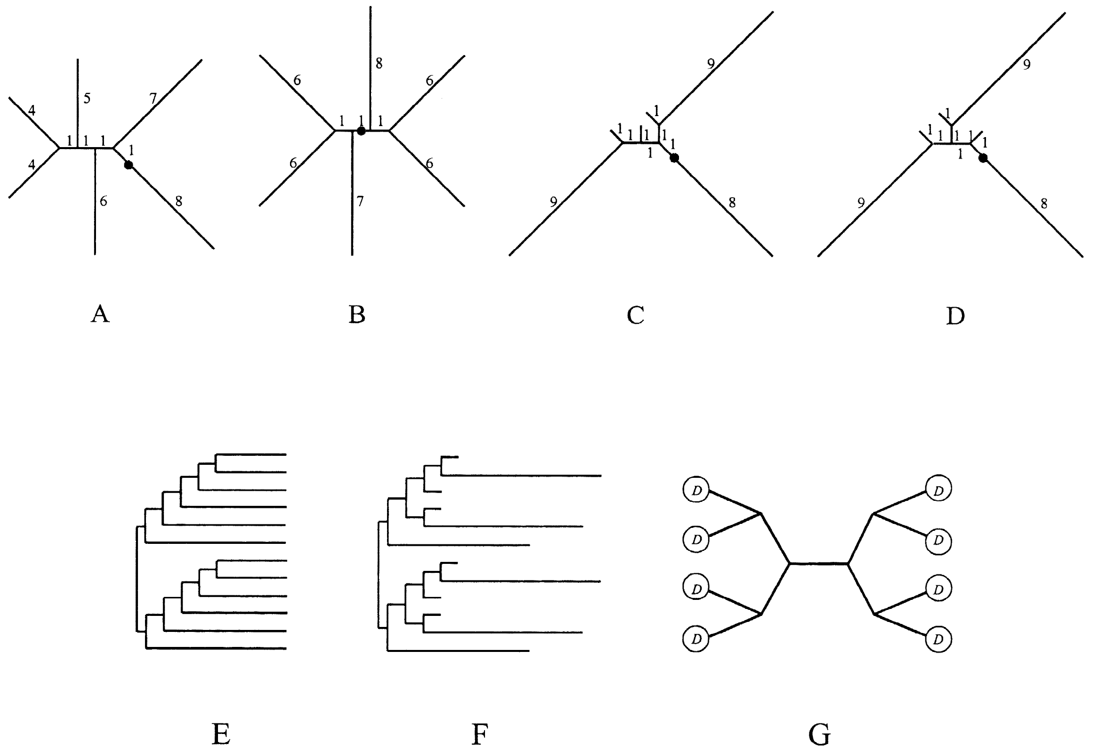

Fig. 1. **A–D.** The four basic model topologies used in this study, with the relative branch lengths shown. Composite trees were constructed by stacking these four trees to give A^x, B^x, C^x, and D^x trees. For example, **E, F,** and **G** are composite trees consisting of two A trees (A^2), two C trees (C^2), and eight D trees (D^8), respectively. All interior branches in the stacked trees have equal relative lengths.

the correct interior branches in the inferred tree (e.g., Hillis 1996; Kim 1998; Nei et al. 1998; Strimmer and von Haeseler 1996). However, a number of specific questions regarding the performance of the NJ method remain unexplored. Are shallow branches (branches closer to the tips of the tree) easier to reconstruct than deeper branches? In this case, shallow branches correspond to more recent evolutionary divergences, whereas deep branches establish evolutionary relationships among groups that have diverged earlier in the evolutionary history. How does an increase in the number of sequences affect the correct inference of shallow and deep branches? What are the relative contributions of the evolutionary rate and sequence length on the efficiency of the NJ method? How long should an interior branch be, in terms of the total number of substitutions, in order to be reconstructed correctly?

Another common feature of previous simulation studies has been that often no distinction was made between expected and realized trees. An expected tree is one in which all branch lengths are expressed in terms of the expected number of nucleotide (or amino acid) substitutions per site, whether or not the evolutionary rate is constant among lineages. A realized tree, on the other hand, has branch lengths equal to the actual number of substitutions per site (Kumar 1996; Nei 1987). The same branch in the realized and expected trees differ in length because evolution is a stochastic process in which the realized tree is one "realization" of the expected tree. The NJ method uses the extant sequences to infer the realized tree rather than the expected tree (Nei and Kumar 2000). Note that the realized tree is not a mere "sample" of the expected tree. Rather, it is an actual quantity to be estimated because the sequences in a real data set are unique products of the evolutionary process which occurs only once for a given gene.

When the expected number of substitutions on a branch is small, the probability that one or more realized branch lengths is equal to zero is high (Kumar 1996). This suggests that for closely related sequences (slowly evolving genes or population level divergence), the topology of the realized tree may contain multifurcations. Therefore, the performance of all tree-making methods should be evaluated by comparing the inferred tree to the realized tree rather than the expected tree (e.g., Kumar 1996; Tateno 1990). What is the difference in the efficiencies of the NJ method in reconstructing the realized versus the expected trees?

It is worth noting that the expected tree can also have branches with expected length equal to zero simply because the product of the evolutionary time elapsed, the length of the gene, and its rate of evolution is practically zero. In such cases, the topology of the expected tree is still bifurcating, but some interior branches are of zero expected length (e.g., Saitou 1996). For simplicity, we have assumed that all interior branches in the model tree have expected branch lengths ≥1 per sequence.

In this paper, we have taken the first step to address

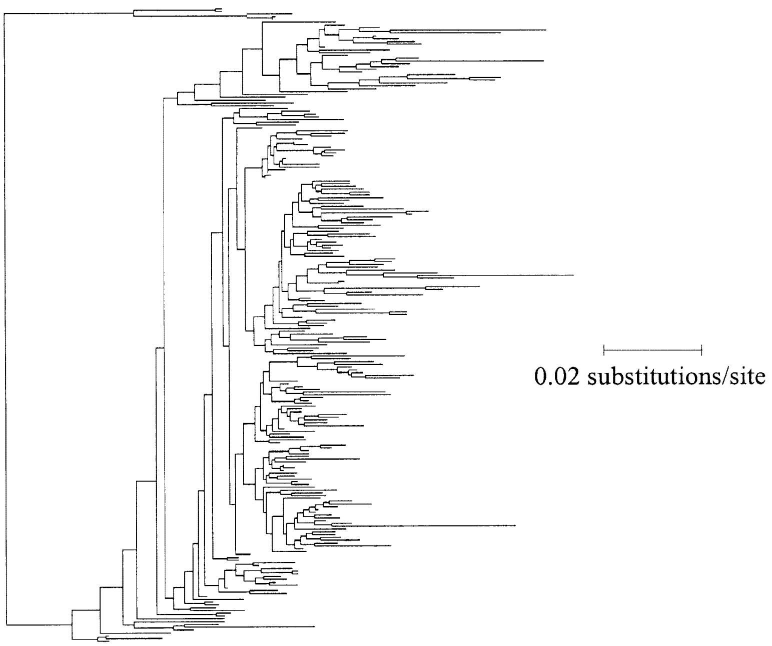

Fig. 2. A 228-taxon *rbc*L tree.

the questions raised above. We discuss the results obtained in relation to the presumed increase in complexity of phylogenetic reconstruction with increasing number of sequences.

Computer Simulations

Model Trees. Following Saitou and Imanishi (1989) and Kumar (1996), we considered four basic six-taxon model trees. These trees are drawn in an unrooted fashion in Figs. 1A–D to reflect the fact that the NJ method produces unrooted trees. Previously these model trees have been drawn with a root (indicated by a filled circle) to specify an arbitrary starting point in the computer simulation. Using these four basic trees, we constructed larger composite phylogenies (Fig. 1), as in Kumar (1996). For instance, Fig. 1E is a composite tree consisting of two copies of tree *A*, where one copy has been grafted onto the other. We refer to topologies generated in this manner as A^x trees, where *x* refers to the number of copies in the composite tree. We constructed A^x, B^x, C^x, and D^x trees, where *x* varied from 1 . . . 10, 16, and 32 (a total of 48 model trees containing up to 192 taxa). In all of these model trees, each interior branch was made to be 1 unit long and the lengths of the external branches are given in multiples of the interior branch length (Figs. 1A–D).

All *expected* interior branch lengths in a given model tree were kept equal in magnitude to compare directly the performance of the NJ method in reconstructing branches at different depths in the tree as a function of the branch location (depth) alone. The stacked tree structure of our large phylogenies also allowed us to study the change in performance of the NJ method from small trees to the larger composite trees. Alternatively, our composite trees can also be viewed as consisting of multiple monophyletic groups, with each group containing the same number of sequences. This situation is similar to that in multigene family evolutionary studies, where gene duplication events need to be inferred and the data are often available for a similar set of model organisms. While our composite trees are convenient for statistical comparisons, the situation in real life is obviously more complicated. Therefore, we also conducted computer simulations using "hybrid" composite trees that were stacked with trees taken at random from among the four basic trees (Figs. 1A–D), as well as a much larger, 228-sequence, chloroplast *rbc*L gene tree (Hillis 1996) containing interior branches of varying expected lengths (Fig. 2) and lacking the repeated phylogenetic structure found in our composite trees. This allowed us to evaluate the generality of the results obtained from the composite trees.

Rates of Evolution and Sequence Length. We conducted computer simulations using many sequence lengths and rates of evolution. Because we are comparing the relative performance of the NJ method in correctly reconstructing small and large phylogenies, we discuss the evolutionary rate in the context of the lengths of interior branches, rather than the maximum pairwise distance between sequences, as the latter depends upon the number of sequences in the data. A low rate of evolution refers to an interior branch length of 0.00625 substitution/site. Multiples of this rate ($r = 0.00625$) were used for all *A–D* model trees as well as the hybrid model trees. For *rbc*L trees, we conducted computer simulations with up to 10-fold rate differences. The sequence lengths employed were in multiples of 100 sites for all the model trees.

Simulating Evolutionary Change. For the computer simulation, the starting point was chosen for each tree (marked by the filled circle in Figs. 1A–D), and for this "root" an ancestral sequence of a given length was first generated by randomly selecting nucleotides such that the four nucleotides are expected to occur with equal frequency in the ancestral sequence. This sequence was evolved by introducing random nucleotide substitutions to generate the immediate descendents. In any given branch, the actual (realized) number of nucleotide substitutions was obtained by selecting a random number from a Poisson distribution with mean equal to the expected number of substitutions (rate × sequence length). A given nucleotide was allowed to change to any of the other three with equal probability, resulting in the Jukes and Cantor (1969) model of nucleotide substitution. This process was carried out for all branches moving away from the root, and a set of sequences was generated at the end of this process. The final set of sequences at the external nodes was then used to reconstruct their evolutionary relationships using the NJ method. We generated 1000 simulation replicates for each case, except for the "hybrid" trees, the 96- and 192-taxon *A–D* trees and the 228-taxon *rbc*L trees, where 100 replications were generated.

Definitions. Tree size refers to the number of sequences. An *interior branch* partitions an unrooted tree into two subtrees, each containing at least two taxa. The *cluster size* for a given interior branch is defined as the minimum of the two subtree sizes. The cluster size thus directly measures the minimum *depth* of a branch in terms of the number of sequences contained in the smaller of the two subtrees that it defines. By this definition, the complexity involved in inferring deep branches is higher than that for shallow branches, because the minimum number of taxa to be joined in inferring deep branches is larger than that for shallow branches in the NJ algorithm. Therefore, the depth of a branch depends only on the subtree sizes rather than the subtree heights in terms of the number of substitutions. This definition of branch depth is more relevant to our analysis because the NJ algorithm always clusters shallow branches before deeper branches, irrespective of the number of substitutions.

Performance measures. A number of different measures were used to quantify the performance of the NJ method.

P_M represents the proportion of all simulation replicates in which the topology of the NJ tree is identical to that of the model tree.

P_{BM} is the proportion of all branches of the model tree that are reconstructed correctly in the NJ tree. $P_{BM} = [c_{avg}/(m - 3)]$, where c_{avg} is the average number of correctly inferred interior branches of the model tree in all simulation replications, and $m - 3$ is the number of interior branches for an unrooted tree containing m sequences.

P_0 is the proportion of branches in the realized tree that receive zero substitutions (zero length branches). $P_0 = [b_{0,avg}/(m - 3)]$, where $b_{0,avg}$ is the average number of zero-length interior branches in the realized tree, in all the simulation replications.

P_{BR} is the proportion of all non-zero-length branches of the realized trees that are reconstructed correctly in the NJ tree. $P_{BR} = c_{>0,avg}/[(m - 3) - b_{0,avg}]$, where $c_{>0,avg}$ is the average number of correctly inferred non-zero-length interior branches in the realized tree, in all simulation replications.

p_B represents the percentage efficiency in correctly estimating branches of a given depth (in terms of the number of taxa) or length (in terms of the number of substitutions). $p_B = b/B$, where B is the total number of occurrences of the desired type of branches (always non-zero length) in all the simulation replicates, and b is the number of cases in which that branch was found in the NJ tree.

Results

Accurate Inference of Complete Trees

Table 1 shows the percentage replicates in which the model tree topology was reconstructed correctly (P_M) for trees containing increasing numbers of sequences and sequence lengths, with $r = 0.0125$. As expected, it is more difficult to reconstruct trees when they contain large numbers of sequences or if the sequences are short (e.g., Kumar 1996; Strimmer and von Haeseler 1996). This is because all $m - 3$ interior branches (nontrivial partitions) need to be reconstructed correctly for correct inference of the complete tree, which requires selecting the sole true tree from a large number of possible trees (Table 2). Longer sequences improve the efficiency of tree-making methods, partly because the pairwise distances can be estimated with better accuracy (lower variance). Table 1 shows slower rates of P_M decline for larger sequences as the number of sequences increases. For instance, for 18 sequences, P_M is 8% for $s = 200$ and 96% for $s = 1000$. When the number of sequences increases to 192, P_M declines to only 46% for $s = 1000$.

Table 1. Percentage replicates in which the complete model tree is reconstructed correctly (P_M) by the NJ method[a]

	Sequence length		
Sequences	200	500	1000
6	57	87	98
12	22	74	96
18	8	63	96
24	3	54	94
30	1	46	93
36	1	39	91
42	0	33	90
48	0	28	89
54	0	25	86
60	0	21	86
96	0	8	79
192	0	0	46

[a] Each value is the arithmetic mean over all the topologies given in Figs. 1A–D, with $r = 0.0125$.

Table 2. Number of unrooted trees and the corresponding number of interior branches for the complete and subtree sizes (numbers of sequences) in the simulation study

Sequences	Unrooted trees	Interior branches
4	3	1
5	15	2
6	105	3
7	945	4
8	10,395	5
9	135,135	6
10	2,027,025	7
12	654,729,075	9
18	$10^{17.28}$	15
24	$10^{26.75}$	21
30	$10^{36.94}$	27
36	$10^{47.69}$	33
42	$10^{58.90}$	39
48	$10^{70.51}$	45
54	$10^{82.45}$	51
60	$10^{94.70}$	57
96	$10^{173.10}$	93
192	$10^{407.79}$	189
228	$10^{502.06}$	225

Influence of Zero-Length Branches on the Efficiency of NJ

Figure 3 shows the mean number of zero-length branches per replication for different tree sizes, with r and s fixed at 0.00625 and 200, respectively. The probability that a given lineage (interior branch) has experienced zero substitutions is given by e^{-b}, where b is the expected branch length in terms of the total number of substitutions per sequence. For $s = 200$ and $r = 0.00625$, $b = 0.00625 \times 200 = 1.25$ substitutions. Since all interior branches

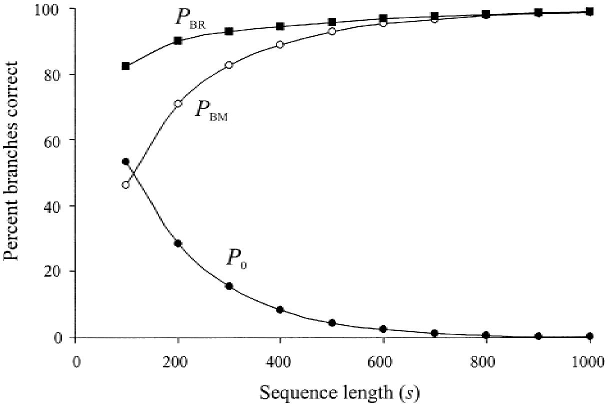

Fig. 3. Percentage branches of the model trees (P_{BM}) and realized trees (P_{BR}) reconstructed correctly by the NJ method, with increasing number of sites, and the corresponding proportion of zero-length branches (P_0; *filled circles*). The values were averaged over all four topologies and all tree sizes (A^x, B^x, C^x, D^x trees), for $r = 0.00625$.

are of equal expected length in our model trees, the expected proportion of zero-length branches is e^{-b}. This expectation is confirmed in the computer simulation results shown in Fig. 3.

Figure 3 also shows the percentage of branches in the model and realized trees that were reconstructed correctly. As expected, the percentage branches correctly inferred increases with increasing sequence length, for model as well as realized trees. However, comparison of the realized tree to the inferred tree shows much higher P_{BR} values even for small s values. Interestingly, P_{BM} (i.e., for the model tree) is essentially a mirror image of P_0, the proportion of zero-length branches in the realized tree. This suggests that the zero-length branches in the realized tree contribute significantly towards the decline in NJ efficiency. Therefore, zero-length branches should be properly discounted in any estimation of the NJ efficiency, as the NJ method reconstructs realized rather than model trees. For this reason, we report only the efficiency of the NJ method in reconstructing non-zero-length branches.

Percentage Branches Reconstructed Correctly

In order to present succinctly the rather voluminous computer simulation results (from the thousands of model trees used) in one place, we first present a summary table (Table 3). In this table, the NJ efficiencies (P_{BR}) were averaged over all rates, topologies, and sequence lengths for a given tree size. Results show that the minimum, maximum, and average NJ efficiencies are similar across tree sizes, which differ 32-fold in the number of sequences (6 to 192). This is further illustrated in Fig. 4, where NJ efficiencies are similar across tree sizes and evolutionary rates, for fixed sequence lengths (100, 200, 500 and 1000). In the figure each value is an average taken from all the topologies for a given tree size and evolutionary rate, for a given sequence length. In general, we find that a medium evolutionary rate leads to a slightly higher performance when compared to lower or higher rates.

Table 3. Percentage branches reconstructed correctly (P_{BR}) for trees of different sizes[a]

	Overall efficiency		
Sequences	Average	Minimum	Maximum
6	92	53	100
12	94	55	100
18	94	54	100
24	95	61	100
30	95	56	100
36	95	60	100
42	95	60	100
48	95	60	100
54	95	60	100
60	95	63	100
96	95	64	100
192	95	61	100

[a] Each value is an average over all rates of evolution ($r = 0.00625$ to 0.0625, in steps of 0.00625), numbers of sites ($s = 100$ to 1000, in steps of 100), and all topologies (A^x, B^x, C^x, and D^x).

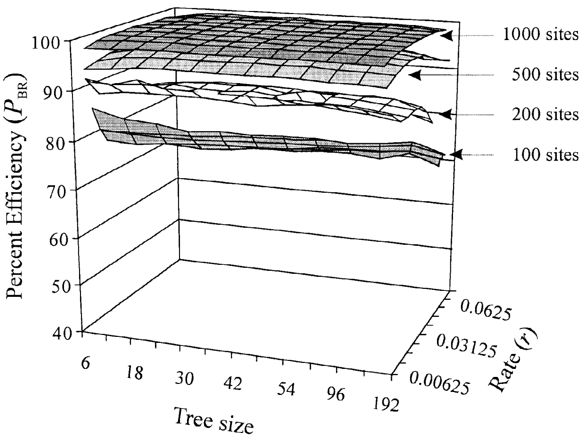

Fig. 4. Percentage efficiency (P_{BR}) of the NJ method for varying evolutionary rates (r) and tree sizes (up to 192 sequences). Average values of P_{BR} from all A^x, B^x, C^x, and D^x trees are shown.

Our large phylogenies consist of four basic trees, each of which constitutes a monophyletic group. Table 4 shows the efficiency with which these monophyletic groups were inferred correctly—observed efficiencies are similar for different tree sizes for a given sequence length. It is thus clear that the NJ method is able to infer groups of the same size with similarly high efficiencies in large as well as small phylogenies.

Effect of Branch Depth on NJ Efficiency (Table 5)

As mentioned earlier, the depth of a branch is defined by the size of the smallest subtree connected to it. Furthermore, a branch is considered correctly inferred when it

Table 4. Percentage replicates in which monophyletic clusters of six taxa were reconstructed correctly[a]

Tree size	Sequence length		
	200	500	1000
12	77	94	99
18	80	96	100
24	83	97	100
30	83	97	100
36	85	97	100
42	84	97	100
48	84	97	100
54	85	97	100
60	85	97	100
96	84	97	100
192	84	96	100

[a] Each value is the arithmetic mean over all the topologies given in Figs. 1A–D and evolutionary rates used.

Table 5. Efficiency of reconstructing branches of various depths[a]

Branch depth	Sequence length		
	200	500	1000
2	86	96	99
3	85	97	100
4	87	97	100
5	86	97	99
6	83	97	100
12	87	99	100
18	89	99	100
24	91	99	100
30	90	99	100
48	92	100	100
96	93	100	100

[a] Each value is a percentage, averaged over all the topologies given in Figs. 1A–D, tree sizes, and evolutionary rates used.

partitions the tree into two clusters, each containing the same set of sequences as in the original tree. In trees with varying *expected* internal branch lengths (e.g., Fig. 2), the efficiency of reconstructing an internal branch could be influenced by the branch depth and/or branch length. This is not the case in our study because all the interior branches in a given tree are of equal expected length. (Of course, the realized interior branch lengths may differ among branches in any given replication). This design allows us to look at only the location (depth) of the branch, independent of its length, and facilitates direct comparison across different parts of a tree. Figure 5 shows the efficiency of reconstruction of branches of various depths for two sequence lengths. For each sequence length we find that the efficiency of reconstruction of interior branches is largely similar across all depths of the tree, with deeper branches in fact being reconstructed with higher efficiency in some cases. This observation is somewhat counterintuitive because deep branches are often thought to be more difficult to reconstruct than the shallow ones (see Discussion later). Furthermore, the efficiency is high for sequence lengths of 500 sites or more and relatively lower for smaller sequences.

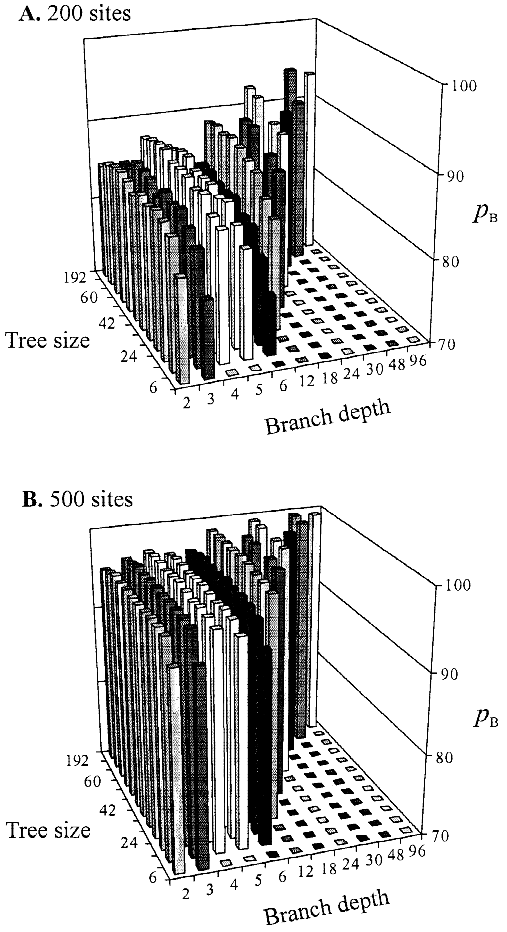

Fig. 5. Probability of correct reconstruction of branches (p_B) at various depths in trees of different sizes. Each p_B value is an average over 10 evolutionary rates and four topologies (A^x, B^x, C^x, and D^x trees).

Efficiency in Reconstructing Branches of Different Realized Lengths

The number of substitutions per sequence that actually occurred in a given branch constitutes the realized length of that branch. This length varies from replication to replication, whereas the expected branch lengths are identical. As mentioned under Computer Simulations, the realized branch lengths are obtained by drawing a random number from a Poisson distribution, with the expected branch length as the mean of the distribution, to simulate the stochastic nature of the evolutionary process. How large should the realized branch length be in order to obtain an NJ efficiency of 95% or higher? Furthermore, how does this length change with sequence length and evolutionary rate (the two determinants of expected branch length)? To address these questions, we computed the percentage efficiency with which branches

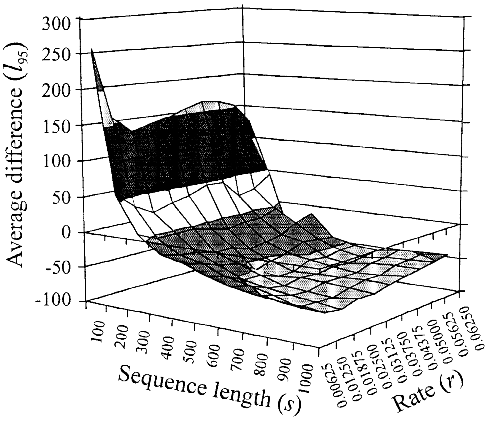

Fig. 6. Average percentage difference between the expected and the minimum realized branch length per sequence needed for $p_B \geq 95\%$.

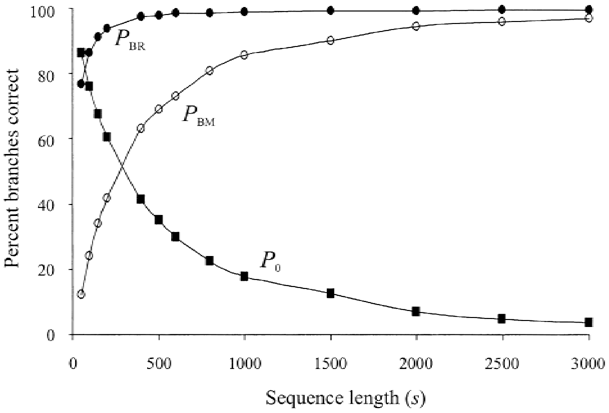

Fig. 7. Percentage branches of the model (P_{BM}) and realized (P_{BR}) *rbc*L trees reconstructed correctly by the NJ method, plotted against increasing number of sites. *Filled circles* show the corresponding proportions of zero-length branches.

of different lengths were constructed correctly, irrespective of their position in the tree. The resulting branch lengths were standardized; $l_{95} = [(b - e)/e] \times 100$, where b is the minimum branch length required for a 95% efficiency, and e is the expected branch length. A negative standardized value shows that a tree with realized branch length that is smaller than the expected branch length can still be reconstructed correctly at an average. This standardization allows us to make comparisons across different rates of evolution and sequence lengths (Fig. 6).

Figure 6 shows that an increase in sequence length (with evolutionary rate held constant) leads to a significant decrease in l_{95}. However, an increase in evolutionary rate (with the sequence length held constant) does not change l_{95}. Therefore, when the total numbers of substitutions in the expected tree are the same, data with longer sequences will perform better than those with faster evolutionary rates. This is not unexpected because the same expected Jukes–Cantor distance will be estimated with lower variance in the former case.

Discussion

In this work, we have presented results from our analysis of large phylogenies in which all interior branches in the expected trees were made equal for any given tree. This stipulation allowed us to examine the relative efficiency with which the deep and shallow interior branches are reconstructed correctly. Furthermore, the stacked structure of our composite model trees is suitable for examining the relative efficiency of correctly reconstructing the same branch in small and large trees. As a result, we are now in a position to establish a "baseline" profile of the NJ performance. In the following we discuss the significance of these results and assess their generality by comparing them to the results obtained from computer simulations involving a 228-taxon *rbc*L tree (Fig. 2) and some "hybrid" composite trees (see Computer Simulations).

Our simulation results clearly establish the adverse effect of the zero-length branches in the realized tree on the performance of the NJ method (Fig. 3). The NJ method is for reconstructing realized trees rather than the expected trees, and therefore, its efficiency should be measured by comparing the inferred tree to the realized tree. For instance, Hillis (1996) conducted a computer simulation using the model tree in Fig. 2 and showed that the NJ method recovers the expected ("model") tree when the sequence length was ~5000 sites. The increase in efficiency of the NJ method in recovering the model tree with increasing number of sites can be attributed to (1) decreasing variance of distance estimates and/or (2) the decrease in the number of zero-length branches. Using the expected branch lengths employed by Hillis (1996), we examined the performance of the NJ method by considering the influence of the zero-length branches (Fig. 7). Our results for the efficiency of the NJ method in reconstructing the model tree are similar to those of Hillis (1996). However, now it is clear that an increase in the sequence length directly reduces the number of zero-length branches, and this is highly correlated (almost as a mirror image) with the efficiency of the NJ method (P_{BM}). In fact, the NJ tree is almost identical to the multifurcating realized tree (>99% branches are correctly inferred) even for only ~500 sites. This result, along with those in Fig. 3, underscores the importance of the distinction between the realized and the model trees in examining the performance of NJ and other methods. In fact, a similar effect is seen when the stepwise addition algorithm is used for the maximum-parsimony method in computer simulations involving the *rbc*L model tree (results not shown). Zero-length branches can be eliminated either by increasing the sequence length or by increasing the evolutionary rate. Our simulations suggest that the former is more effective than the latter, as the distances can be estimated with lower variances in the former case (also see Fig. 6).

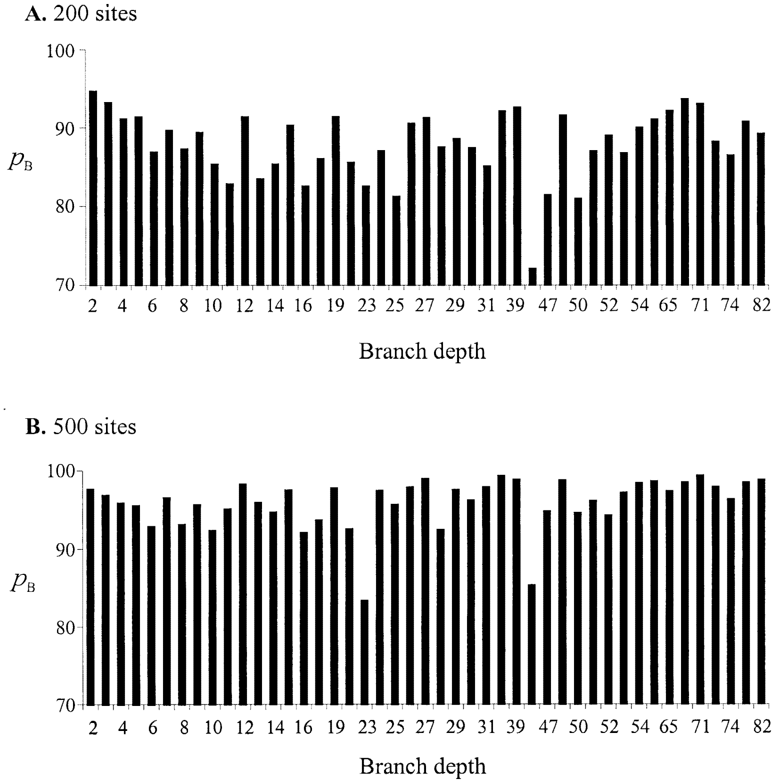

Fig. 8. Reconstruction efficiency for branches at different depths in the *rbc*L tree, for 200 sites **(A)** and 500 sites **(B).**

The efficiency of NJ in terms of the proportion of branches reconstructed correctly (P_{BR}) in the realized trees is similar for trees consisting of vastly different number of sequences. Strimmer and von Haeseler (1996) have reported similar results, but they did not remove the negative contribution made by zero-length branches. Since P_{BR} is similar for large and small phylogenies (Table 3), the efficiency of reconstructing deeper branches is likely to be no worse than that of the shallow branches, if their expected branch lengths are equal. This was indeed the case, as the branches at different depths are inferred correctly with similar efficiencies (Fig. 5). In fact, deeper branches appear to be reconstructed correctly with a higher probability, in some cases. This is due in part to the fact that the estimate of the average distance between groups of sequences, used in reconstructing deep branches, has a lower variance.

An extrapolation of the results in Fig. 5 comes from the analysis of our “hybrid” trees as well as the unequal-internal branch length *rbc*L tree, which contains branches with depths ranging from 2 to 82 taxa and of different lengths. The reconstruction efficiency remained similar across different branch depths, as long as only the non zero-length interior branches in the realized trees were considered for measuring the efficiency (Fig. 8). Recently, many investigators have considered the effect of taxon sampling on the efficiency of tree-making methods, in which the main emphasis has been to study and remedy the effect of long branch attraction for small and large phylogenies (see Graybeal 1998; Hillis 1998; Kim 1996; Purvis and Quicke 1997; Yang and Goldman 1997). The large model trees used in our computer simulations were formed by stacking smaller trees. This increases the tree size by the addition of sister groups to existing clusters, rather than the addition of taxa to break up long branches, as done in taxon-sampling studies. Therefore, a comparison of our results to those by above authors is not straightforward.

The NJ method works in a stepwise fashion, inferring shallow branches first. Therefore, the topological errors in the early stages of tree reconstruction may propagate as we move toward inferring deeper branches. Consequently, one may expect deep branches to be more difficult to reconstruct correctly than shallow branches, all else being equal. However, this intuitive argument is clearly not supported, as the efficiency does not decline with depth (Figs. 5 and 8), suggesting that the accuracy of the NJ method in the later stages of clustering (deep branches) is largely independent of the accuracy at the early stages (shallow branches). To look for an explanation of why this may happen, let us consider the theoretical aspects of the minimum-evolution (ME) principle (Rzhetsky et al. 1995; Rzhetsky and Nei 1992) that forms the basis of the NJ method.

Consider the tree in Fig. 9, where groups *I* and *J* are neighbors, as are groups *K* and *L*. If groups *I, J, K,* and *L* are reconstructed correctly, then, under the ME principle, the correct inference of branch *e* is not affected by inaccuracies in inferring within-group phylogenies (Rzhetsky et al. 1995). That is, errors in phylogeny within groups do not affect higher-level clustering as long as the monophyly of a group is inferred correctly,

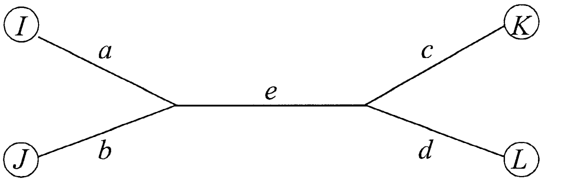

Fig. 9. A schematic showing the topological configuration around an interior branch (*e*). Four clusters (*I, J, K,* and *L*) always surround any given internal branch.

and the realized branch length (branch *e*) will dictate the efficiency of correct reconstruction of that branch. When monophyletic relationships within groups (*I, J, K,* and *L*) are not correctly inferred, then one of two things may happen. First, group *I* may contain some taxa that belong to group *J,* and vice versa (or such swapping may occur for *K* and *L*). In this case, the reconstruction of *e* may not be affected because the true neighbor groups will tend to cluster together anyway. The second possibility is the incorrect grouping of taxa from more distantly related clusters (e.g., taxa from group *I* clustering within group *K*). This would depend on the length of branch *e:* for longer *e* it is more difficult for a taxon to cross over to a nonsister group. For a deep branch, the number of taxa around it is large and the random possibility of crossover of one or more taxa is potentially larger than for a shallow branch. However computer simulations for deep branches show that this is not the case. Rather, the efficiency is at least the same, or sometimes greater, which, as mentioned earlier, is because the average distances between groups have lower variance than the pairwise distances for individual sequences.

In maximum-parsimony analysis, it is generally thought that homoplasy will hinder accurate reconstruction of higher-level relationships (deep branches), as the phylogenetic signal to infer deep branches may deteriorate with later evolutionary changes. This intuition needs to be examined by computer simulation. In the case of distance-matrix methods such as the NJ method, however, pairwise distances are computed in a step independent of the reconstruction of the evolutionary histories, and, perhaps consequently, the degradation of the phylogenetic signal does not appear to occur.

The argument presented above is only approximate, as the NJ method implements ME criterion locally (at each stage of clustering) rather than globally. We have examined the performance of NJ with respect to the same optimality criterion as for ME. That is, we compared the sum of ordinary least-squares estimates of branch lengths (*S*) of each of our NJ trees with that of the corresponding model trees. We found that the NJ tree is less optimal than the true tree in only 22% of the replicates (see also Nei et al. 1998). In those cases, the least optimal NJ tree was only 3% worse than the true tree in terms of the *S* value. Relating the effects of sequence length and evolutionary rate on the percentage optimality score difference (Nei et al. 1998) between the NJ and the model trees (not shown), we found that for a given sequence length there is little difference between the optimality score for the NJ topology and that for the model topology, irrespective of the tree size. The worst NJ performance was for large trees with very small sequence lengths (or very slow evolutionary rates), as there was a very large number of statistically equally good trees (Kumar 1996). Therefore, the results presented in this paper are generally applicable to methods with underlying principles similar to the NJ method (e.g., Gascuel 1997).

It is important to exercise caution in extrapolating results from any computer simulation to real-life situations. We have assumed that the evolutionary processes among the lineages have remained the same throughout the evolutionary history, i.e., the evolutionary process is stationary. This condition is often met in short-term evolution (e.g., population data) and in slowly evolving genes but is likely to be violated when we consider long-term evolutionary histories of genes and species. If the stationarity condition is not met, the correct inference of deep branches is likely to be adversely affected (e.g., Steel et al. 1993). This aspect will be examined in further computer simulation studies.

Acknowledgments. We would like to thank Roman G. Johnson and Tushar S. Gadagkar for helping with the simulations and Dr. Thomas E. Dowling for his comments on the manuscript. We thank two anonymous reviewers for their helpful comments. This research was supported by a National Institutes of Health research grant to S.K.

References

Balczarek K, Lai Z-C, Kumar S (1997) Evolution and functional diversification of paired box (*Pax*) DNA-binding domains. Mol Biol Evol 14:829–842

Charleston MA, Hendy MD, Penny D (1993) Neighbor-joining uses the optimal weight for net divergence. Mol Phylogenet Evol 2:6–12

Duret L, Mouchiroud D, Gouy M (1994) HOVERGEN: A database of homologous vertebrate genes. Nucleic Acids Res 22:2360–2365

Gascuel O (1994) A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. Mol Biol Evol 11:961–963

Gascuel O (1997) Concerning the NJ algorithm and its unweighted version, UNJ. In: Mirkin B, McMorris FR, Roberts FS, Rzhetsky A (eds) Mathematical hierarchies and biology. American Mathematical Society, Providence, RI, pp 149–170

Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol 47:9–17

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. In: Doolittle RF (ed) Methods in enzymology. Academic Press, San Diego, pp 383–401

Hillis DM (1996) Inferring complex phylogenies. Nature 383:130–131

Hillis DM (1998) Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol 47:3–8

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kim J (1996) General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst Biol 45:363–374

Kim J (1998) Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst Biol 47:43–60

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Kumar S (1996) A stepwise algorithm for finding minimum evolution trees. Mol Biol Evol 13:584–593

Kumar S, Rzhetsky A (1996) Evolutionary relationships of eukaryotic kingdoms. J Mol Evol 42:183–193

Kumar S, Balczarek KA, Lai Z-C (1996) Evolution of the *hedgehog* gene family. Genetics 142:965–972

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York

Nei M, Kumar S, Takahashi K (1998) The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci USA 95:12390–12397

Purvis A, Quicke DLJ (1997) Building phylogenies: Are the big easy? Trends Ecol Evol 12:49–50

Rzhetsky A, Nei M (1992) Statistical properties of the ordinary least-squares, generalized least-squares and minimum-evolution methods of phylogenetic inference. J Mol Evol 35:367–375

Rzhetsky A, Kumar S, Nei M (1995) Four cluster analysis: A simple method to test phylogenetic hypotheses. Mol Biol Evol 12:163–167

Saitou N (1996) Reconstruction of gene trees from sequence data. Methods Enzymol 266:427–449

Saitou N, Imanishi M (1989) Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree reconstruction in obtaining the correct tree. Mol Biol Evol 6:514–525

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Steel MA, Lockhart PJ, Penny D (1993) Confidence in evolutionary trees from biological sequence data. Nature 364:440–442

Strimmer K, von Haeseler A (1996) Accuracy of neighbor joining for n-taxon trees. Syst Biol 45:516–523

Tateno Y (1990) A method for molecular phylogeny construction by direct use of nucleotide sequence data. J Mol Evol 30:85–93

Tateno Y, Takezaki N, Nei M (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

Yang Z, Goldman N (1997) Are big trees indeed easy? Trends Ecol Evol 12:357