Prospects for inferring very large phylogenies by using the neighbor-joining method

Koichiro Tamura*,†, Masatoshi Nei‡, and Sudhir Kumar*§

*Center for Evolutionary Functional Genomics, The Biodesign Institute, and †School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501; ‡Department of Biological Sciences, Tokyo Metropolitan University, Tokyo 192-0397, Japan; and §Department of Biology, Pennsylvania State University, University Park, PA 16802

Contributed by Masatoshi Nei, June 11, 2004

Current efforts to reconstruct the tree of life and histories of multigene families demand the inference of phylogenies consisting of thousands of gene sequences. However, for such large data sets even a moderate exploration of the tree space needed to identify the optimal tree is virtually impossible. For these cases the neighbor-joining (NJ) method is frequently used because of its demonstrated accuracy for smaller data sets and its computational speed. As data sets grow, however, the fraction of the tree space examined by the NJ algorithm becomes minuscule. Here, we report the results of our computer simulation for examining the accuracy of NJ trees for inferring very large phylogenies. First we present a likelihood method for the simultaneous estimation of all pairwise distances by using biologically realistic models of nucleotide substitution. Use of this method corrects up to 60% of NJ tree errors. Our simulation results show that the accuracy of NJ trees decline only by ~5% when the number of sequences used increases from 32 to 4,096 (128 times) even in the presence of extensive variation in the evolutionary rate among lineages or significant biases in the nucleotide composition and transition/transversion ratio. Our results encourage the use of complex models of nucleotide substitution for estimating evolutionary distances and hint at bright prospects for the application of the NJ and related methods in inferring large phylogenies.

The NJ method constructs trees by clustering neighboring sequences in a stepwise manner. In each step of sequence clustering, it minimizes the sum of branch lengths (2) and thus examines multiple topologies. For large data sets, however, NJ examines only a minuscule fraction of the total number of possible topologies. For instance, it will examine all three unrooted trees in the case of four sequences but only 1010 of >1013,867 possible trees when 4,000 sequences are used (Fig. 1B).

The effect of this property on the accuracy of NJ trees has been unclear. Therefore, we have done computer simulations to study the accuracy of NJ trees when tens to thousands of sequences are used.

NJ is known to be statistically consistent in the sense that, if correct pairwise distances with no statistical errors are used, it reconstructs the true tree (2). In practice, however, estimates of all distances are subject to statistical errors, so it may produce erroneous trees. At present, all distances are estimated independently for each pair of sequences [independent estimation (IE) method] either by analytical formulas (1, 14, 15) or by likelihood methods (4, 15). The standard errors of the estimates obtained in this way are rather high unless very long sequences are used. However, these standard errors can be reduced considerably if all distances for a given set of aligned sequences are estimated simultaneously [simultaneous estimation (SE) method]. Recently, we developed an SE method based on the maximum likelihood principle and found that the use of this method substantially improves the accuracy of NJ trees. In this article, we first present this SE method and then discuss the accuracy of NJ trees obtained for large and small phylogenies.

Methods

Simultaneously Estimating All Pairwise Distances. Pairwise distances used for constructing NJ trees are currently estimated by the IE method for a variety of mathematical models to incorporate varying degrees of complexity of nucleotide or amino acid substitution (reviewed in refs. 1, 4, 14, 15). The estimates obtained by the IE method are expected to have larger standard errors than those obtained by the SE method. These larger errors are partly because the parameters, such as the transition/transversion rate ratio in the Kimura model (16), are estimated independently for each pair of sequences in the IE method, whereas they are estimated by considering all pairs of sequences simultaneously in the SE method. There is obviously a greater advantage of using the SE method for a sophisticated model of substitution than for a simple model, because in the former model we have to estimate a larger number of parameters. Furthermore, the accuracy of parameter estimates obtained by the SE method is expected to increase as the number of sequences increases.

In the following we consider the evolutionary distance based on the TN model (3), which is one of the most sophisticated models of nucleotide substitution. For this model, the evolutionary distance (d) is given by

\[
d = 4(g_{a,C}a_1 + g_{a,G}a_2 + g_{a,T}a_3),
\]

where \(a_1, a_2, \text{ and } b\) are the numbers of transitional substitutions between purines \((a_1)\), transitional substitutions between pyrimidines \((a_2)\), and transversional substitutions \((b)\) per site, respectively.
the total number of topologies examined is given by $g$ that are examined by the neighbor-joining (NJ) method in producing a final bijection of data sets in which unestimable distances occurred was computed. For each model tree, 100 data sets were generated, and the proportion of data sets in which the original Tamura–Nei distance was not applicable for one or more pairwise comparisons for the HKY distance for sequences $i$ and $j$ are compared, respectively, $P_{ij}$, $P_{2ij}$, and $Q_{ij}$ are the theoretical values of $P_{ij}$, $P_{2ij}$, and $Q_{ij}$, respectively, and are given by

$$P_{ij} = \frac{2g_AG}{g_R} \left[ g_R \exp(-2g_A k_i + g_Y b_{ij}) + g_Y \exp(-2b_{ij}) \right],$$

$$P_{2ij} = \frac{2g_AG}{g_Y} \left[ g_Y \exp(-2g_Y k_i + 2g_R b_{ij}) + g_R \exp(-2b_{ij}) \right],$$

$$Q_{ij} = 2g_g [1 - \exp(-2b_{ij})],$$

where $g_R = g_A + g_G$ and $g_Y = g_C + g_T$.

However, because the parameters $k_1$ and $k_2$ are shared by the log likelihood functions for all pairs of $i$ and $j$, they should be estimated by maximizing the sum of all likelihood functions (SL), which is

$$SL = \sum_{i} \sum_{j<i} L_{ij}.$$  

Theoretically, this is not a likelihood function, because $d_{ij}$'s are not necessarily independent. However, by using the method of Taylor's expansion (as in ref. 17), one can show that approximately unbiased estimates of $k_1$ and $k_2$ can be obtained by maximizing SL. We therefore suggest the following procedure for estimating $k_1$, $k_2$, and $b_{ij}$'s. We first compute the initial estimates of $k_1$, $k_2$, and $b_{ij}$'s by using uncorrected estimates, $k_1 = (P_{ij}/g_A g_Y)/(Q_{ij}/g_A g_Y)$, $k_2 = (P_{2ij}/g_Y g_C)/(Q_{ij}/g_A g_Y)$, and $b_{ij} = Q_{ij}/(4g_g g_Y)$. This method is computationally efficient and gives estimates with smaller standard errors. We then compute the averages of estimates of $k_1$, $k_2$, and $b_{ij}$'s separately and use them for obtaining improved estimates of $b_{ij}$'s by maximizing $L_{ij}$ in Eq. 3. We can now obtain improved estimates of $k_1$, $k_2$, and SL by maximizing SL in Eq. 7. These estimates are then used for further improvement of the estimates of $b_{ij}$'s by Eq. 3. The last two processes are repeated until stable estimates of $k_1$, $k_2$, and $b_{ij}$'s are obtained. The final estimates are asymptotically unbiased, although they may not be maximum likelihood estimates.

The SE method above can be used for many other substitution models. For example, the Hasegawa–Kishino–Yano (HKY) model (18) is a special case of the TN model, in which $k_1$ and $k_2$ are assumed to be the same. Therefore, the HKY distance for sequences $i$ and $j$ can be estimated by using Eq. 3 under the assumption of $k_1 = k_2 = k$. Similarly, the Tamura (19) and the Kimura (16) models are a special case of the TN model (see ref. 1). In the case of the Tamura model the $G + C$ content ($\theta = g_C + g_G$) instead of nucleotide frequencies is considered with a single $k$ parameter. In the Kimura model all nucleotide frequencies are assumed to be equal (0.25) with a single $k$. The SE method can also be used when the substitution rate varies with site, following the gamma or other distribution (e.g., refs. 20 and 21).
In Eqs. 4–6, the use of average nucleotide frequencies for the entire set of sequences is recommended because of the smaller sampling errors. However, when the distance methods that relax the assumption of the equality of nucleotide frequencies among lineages (heterogeneity of substitution pattern over the phylogeny) are used, the sequence-specific nucleotide frequencies should be used for each comparison (22). Note that these methods do not take into account the variation of transition/transversion ratio among lineages and are not guaranteed to generate asymptotically unbiased estimates of evolutionary distances.

The SE approach easily produces estimates of the transition/transversion rate ratios \( k_1 \) and \( k_2 \) and pairwise distances \( d_{ij} \) for all sequence pairs simultaneously. Computation of the variances of these estimates by analytical formulas is somewhat complicated, but they can be obtained by the bootstrap method with site resampling (1, 23).

**Methods of Computer Simulation.** In our study of the accuracy of NJ trees obtained by the IE and SE methods of distance estimation, we used three different sets of model trees. The first model tree consisted of 66 sequences (Fig. 2A), of which the relative branch lengths were derived from the mammalian DNA sequence data (figure 1 in ref. 24). For this model tree, we considered 448 different sets of evolutionary parameters (substitution parameters and sequence lengths), in which the number of nucleotides per sequence \( n \) varied from 147 to 9,359, the evolutionary rate varied 10 times, the G + C content \( (\theta) \) varied from 0.3 to 0.9, and the transition/transversion rate ratio \( k_1 = k_2 = k \) varied from 2.1 to 26.6 (see ref. 25). Using this model tree and the set of evolutionary parameters with the TN model, we evaluated the relative performance of the SE and IE methods in estimating evolutionary parameters and inferring phylogenies by the NJ method.

The second and third sets of model trees were used to compare the accuracy of NJ trees of different sizes. The second set was based on two predefined model trees in Fig. 2, where the tree in \( B \) represents a case of constant rate evolution for 32 sequences and the tree in \( C \) represents a varying rate case with 32 sequences. Multiple copies of these trees, connected at the roots, generated increasingly larger model trees consisting of 32 to 4,096 sequences (see also ref. 10). The third set of model trees was generated by using an agglomerative algorithm. In this algorithm we combined a given set of sequences and randomly selected pairs or groups of sequences for making different model trees in a stepwise manner. This process was continued until the required number of sequences was clustered. Model trees in Fig. 2D and E are two examples of such randomly generated trees. For model trees in Fig. 2B and D, the exterior branch lengths (0.2 substitutions per site) were 20 times longer than the interior branch lengths (0.01 substitutions per site). For model trees in Fig. 2C and E, the exterior branches were either long (0.2) or short (0.01). For the type of model trees in Fig. 2E, the exterior branches were assigned a long or short length randomly. The number of nucleotides per sequence \( n \) used in the second and third data sets was 1,000.

For the second and third sets of model trees, nucleotide substitution was simulated by using the TN model with two sets of biologically realistic values of substitution parameters: \( i ) \ k_1 = k_2 = 4 \) and \( g_A = g_T = g_C = g_G = 0.25 \) to simulate nuclear gene evolution and \( ii ) \ k_1 = k_2 = 20, g_A = g_T = 0.40, \) and \( g_C = g_G = 0.10 \) to simulate animal mitochondrial DNA gene evolution.

Using the standard simulation procedure (e.g., refs. 1 and 10), we generated simulated sequence data for each model tree with a set of nucleotide substitution parameters and sequence length \( n \). A NJ tree was then constructed and compared with the true tree. The accuracy of the NJ tree was measured by the percentage of phylogenetic clades correctly inferred \( (P_C) \). This was obtained by \( P_C = 100 \left( 1 - d_T / (2m - 6) \right) \), where \( d_T \) is the topological distance between the inferred and model trees and \( m \) is the number of sequences used \( (1, 26, 27) \). \( P_C \) is 100% when all clades are correctly inferred and is 0% when none of the clades is correctly identified. For each model tree (model tree in Fig. 2A, B, or C) for a given number of sequences, 100 replicates of simulated sequence data were generated for the same topol-
ratios was used for generating sequence data with the transition
the SE method. (Generated sequences according to the model tree in Fig. 2
distances obtained by the IE and SE methods for a data set of 66 computer-
tained by the IE and SE methods. Fig. 3
Tamura et al
Performance of the SE Method.
Fig. 2
D
Fig. 3. Standard errors and transition/transversion rate ratios obtained by
the SE method. (A) Relationships between the standard errors of evolutionary
distances obtained by the IE and SE methods for a data set of 66 computer-
tained by the IE and SE methods. Fig. 3
E
nucleotides. There are (66 × 65)/2 = 2,145 data points in this figure. (B) Relationships between the estimated and true values of k (= [k1 + k2]/2) for
448 different patterns of nucleotide substitutions (see text). For each pattern
of nucleotide substitution, the average value (circles) and the 95% confidence
limits (lines) of the estimate of k were obtained from 100 data sets generated
by computer simulation with the model tree in Fig. 2A.

To make the PC values comparable among model trees of
different sizes (except Fig. 2A), the expected interior branch
lengths were assumed to be the same (0.01 substitutions per site)
for all topologies. This approach is different from that used in
most previous studies, in which the maximum depths of trees or
the maximum evolutionary distances were assumed to be the
same (e.g., refs. 11 and 28). The latter approach makes the
interior branch lengths shorter in large phylogenies than in small
phylogenies and, therefore, the comparison of PC values among
trees of different sizes becomes improper.

Results
Performance of the SE Method. Using the model tree in Fig. 2A, we
first compared the standard errors of distance estimates ob-
tained by the IE and SE methods. Fig. 3A shows that the standard
distances (proportions of different nucleotide
sites) are used than when IE distance estimates for the correct
substitution models are used (1, 11, 29, 30). This occurred mainly
because p distances have smaller standard errors than distances
estimated by the IE method. The same results were observed in

Fig. 4. Accuracy of NJ-SE trees. (A) Accuracies (PC) of NJ trees when the SE and
IE methods with the TN model were used for the 66-sequence model tree in
Fig. 2A (see text). (B) Comparison of PC values for NJ-SE trees with those for NJ
trees obtained with p-distance (NJ-p). For each simulation condition, average
PC values from 100 replicate simulations are plotted.
For mitochondrial gene evolution, \( \text{fi} \) were used for both type D trees \( (\text{fi}^1) \) when the TN model was used. However, NJ-SE trees had a higher accuracy of \( 60.4\% \) for the simulation conditions when the TN model was used. For nuclear gene evolution, \( \text{fi} \) was used. When the evolution of mitochondrial DNA was considered, \( \text{PC} \) was much lower than that for nuclear genes \( (55\% \text{ for } m = 32) \) \( (\text{fi}^1) \). Yet, the decline of \( \text{PC} \) with increasing \( m \) was quite small \( (\text{fi}^1) \). The \( \text{PC} \) value for \( m = 4,096 \) was smaller than that for \( m = 32 \) only by 2.2\%. A small extent of decline of \( \text{PC} \) with increasing \( m \) was observed even when type C model trees \( (\text{fi}^1) \) were used \( (\text{fi}^1) \) or when randomly generated model trees \( (\text{fi}^1) \) \( (\text{fi}^1) \) were used \( (\text{fi}^1) \). The decrease in \( \text{PC} \) was \( <10\% \) in every case in \( (\text{fi}^1) \). These results indicate that \( \text{PC} \) does not decline very much when \( m \) increases from 32 to 4,096 and, therefore, \( \text{NJ} \) can be used efficiently even when hundreds or thousands of sequences are used.

Similar results were obtained for NJ-SE trees when the substitution pattern and \( G + C \) content vary with evolutionary lineage or when the sequence length is reduced to a half \( (\text{fi}^1) \) (results not shown). Therefore, although the absolute values of \( \text{PC} \) depend on the substitution pattern or sequence length, the relationship between \( \text{PC} \) and \( m \) is nearly the same for all cases. In other words, \( \text{PC} \) generally declines with increasing \( m \), but the extent of the decline is remarkably small.

**Discussion**

We have seen that the SE of evolutionary distances reduces the variances of distance estimates considerably and consequently increases the accuracy \( (\text{PC}) \) of NJ trees. We have also seen that the \( \text{PC} \) of NJ trees does not decrease significantly when the number of sequences \( (m) \) increases from 32 to 4,096. However, this latter property is not necessarily due to the use of SE distances, because a similar property was observed even with IE distances when \( m \) was \( \leq 100 \) \( (\text{fi}^1) \). Some authors have reported a substantial decrease of \( \text{PC} \) when \( m \) increased from 50 to 100 \( (\text{fi}^1) \) \( (\text{fi}^1) \) \( (\text{fi}^1) \). In the latter study the lengths of interior branches were considerably smaller for large trees than for small trees, and this difference contributed to the substantial decrease of \( \text{PC} \). Therefore, it is important to maintain the same interior branch lengths for a comparison of \( \text{PC} \) for large and small trees, as mentioned earlier.
In reality, however, our assumption that all interior branches have the same length irrespective of the tree size is unrealistic. When the number of sequences used is very large, it is quite likely that some interior branches are relatively long and others are very short. If the expected length of an interior branch is very short, no substitutions may occur in the branch when the number of nucleotides examined is small. In this case it would be difficult to resolve the clades associated with the short branches. Therefore, our results should be interpreted only as a guideline. As long as the interior branch is sufficiently long and a sufficient number of nucleotides per sequence is used, the NJ method is capable of producing reasonably accurate trees.

The relatively high $P_C$ values obtained for large trees in this study are partly due to the use of SE distances, which have smaller standard errors than IE distances. This SE approach is effectively the same as the use of a larger number of nucleotides in the IE approach. For example, when model tree $B$ was used, the average standard error of SE distances for the TN model with $k_1 = k_2 = 20$, $\mu_A = \mu_T = 0.4$, $g_c = gc = 0.1$, and $n = 1,000$ was similar to that of IE distances with $n = 1,650$. Furthermore, another merit of using the SE approach instead of the IE approach exists. That is, when the number of sequences ($m$) is as large as 1,000, estimates of IE distances are not always obtainable, because the arguments of logarithms in the analytical formulas may become negative by chance (31). The proportion of such unestimable cases increases as $m$ increases in the IE method (Fig. 1C). By contrast, we have encountered no such cases in our simulation when the SE method was used. Of course, unestimable cases might happen even with the SE method, if the extent of sequence divergence is very high, but the probability of occurrence of such events is much smaller in the SE approach than in the IE approach.

The higher accuracy of NJ-SE trees also appears to be related to the fact that the sum of branch lengths ($S$) for NJ-SE trees is on average closer to that of the true tree than the $S$ value for NJ-IE trees (Fig. 6). It is already known that the $S$ value for NJ-IE trees is on average considerably smaller than that of the true tree when the number of sequences relative to the sequence length is small (8, 11, 12, 32). Our simulation results indicate that the $S$ values for NJ-SE trees are closer to that of the true tree although, in general, they are still smaller than that of the true tree (Fig. 6). These results also indicate that although the NJ method examines a minuscule portion of the entire tree space, a further search for trees with smaller $S$ values does not necessarily lead to a tree closer to the true tree. This is because the optimization principle does not work well in phylogenetic inference when $m$ is large relative to $n$, whether the minimum evolution, maximum parsimony, or maximum likelihood method is used (1, 32).

We therefore conclude that the prospects are bright for using the NJ method for generating initial evolutionary hypotheses for very large species and multigene family trees. These large phylogenies often span long evolutionary times in which the pattern of nucleotide substitution is expected to be complex (1, 4). In these cases, the use of the SE approach with sophisticated models of DNA substitution is expected to improve the accuracy of phylogenetic inference.

We thank C. R. Rao, Xun Gu, George Zhang, Adriana Briscoe, Alan Filipski, Araxi Urritia, Dana Desonic, Michael Rosenberg, Sankar Subramanian, and Sudhindra Gadagkar for comments. This work was supported by grants from the National Institutes of Health (to S.K. and M.N.), the National Science Foundation (to S.K. and James L. Collins, School of Life Sciences, Arizona State University), and the Japan Society for the Promotion of Science (to K.T.).