Generalized Neighbor-Joining: More Reliable Phylogenetic Tree Reconstruction

William R. Pearson,† Gabriel Robins, and Tongtong Zhang**

*Department of Computer Science and †Department of Biochemistry, University of Virginia

We have developed a phylogenetic tree reconstruction method that detects and reports multiple topologically distant low-cost solutions. Our method is a generalization of the neighbor-joining method of Saitou and Nei and affords a more thorough sampling of the solution space by keeping track of multiple partial solutions during its execution. The scope of the solution space sampling is controlled by a pair of user-specified parameters—the total number of alternate solutions and the number of alternate solutions that are randomly selected—effecting a smooth trade-off between run time and solution quality and diversity. This method can discover topologically distinct low-cost solutions. In tests on biological and synthetic data sets using either the least-squares distance or minimum-evolution criterion, the method consistently performed as well as, or better than, both the neighbor-joining heuristic and the PHYLIP implementation of the Fitch-Margoliash distance measure. In addition, the method identified alternative tree topologies with costs within 1% or 2% of the best, but with topological distances of 9 or more partitions from the best solution (16 taxa); with 32 taxa, topologies were obtained 17 (least-squares) and 22 (minimum-evolution) partitions from the best topology when 200 partial solutions were retained. Thus, the method can find lower-cost tree topologies and near-best tree topologies that are significantly different from the best topology.

Introduction

Reconstruction of ancestral relationships from contemporary data is widely used to provide both evolutionary and functional insights into biological systems. The explosive increase in available DNA sequence data has increased interest in phylogenetic analysis of multigene and domain-swapped protein families. Three general classes of phylogenetic reconstruction methods are commonly used for analysis of sequence data sets: parsimony methods (Swofford et al. 1996), distance-based methods (Fitch and Margoliash 1967), and maximum-likelihood methods (Felsentein 1982; 1988). Parsimony- and distance-based methods are most often used, largely because they are faster computationally and allow a larger number of potential phylogenetic trees to be evaluated.

Reconstruction of an evolutionary history for a set of contemporary taxa based on their pairwise distance is computationally intractable (i.e., NP-complete) for various optimality criteria (Foulds and Graham 1982; Day 1987), including the least-squares criterion and the minimum-evolution criterion. Various heuristics have been proposed to search for solutions of desired quality (Felsenstein 1988; Bandelt and Dress 1992; Swofford et al. 1996), and the majority of these methods are greedy methods, which always employ moves that are ''locally best'' and may not necessarily lead to global optima (Swofford et al. 1996). Among the greedy approaches, the neighbor-joining (NJ) method (Saitou and Nei 1987; Studier and Keppler 1988) is widely used by molecular biologists due to its efficiency and simplicity.

Greedy methods are efficient because they explore only a small portion of the solution space. (A solution space is the set of all possible phylogenies spanning the given taxa.) Taxa correspond to leaves in a tree that spans them. However, greedy methods can fail to find the best overall solution if they become ''trapped'' in local optima. In addition, because only a small fraction of the solution space is examined, a greedy heuristic typically will not report (or detect) alternative solutions with distinct topologies that may fit the data nearly as well, or even equally well. Neglecting such alternative solutions can produce misleading inferences regarding evolutionary history. For instance, Wilson et al. (1989) concluded that all humans originated from Africa, because their tree-building method failed to discover alternative, near-optimal trees that were consistent with a different geographical history (Maddison 1991).

To improve the reliability of phylogenetic tree reconstruction, we propose a scheme which samples the solution space more extensively by repeatedly using the NJ algorithm (Saitou and Nei 1987). Instead of tracking only a single locally best tree as NJ does, our scheme maintains multiple partial solutions as it progresses. The method explores all possible trees derivable from the set of current partial solutions in a single NJ step, and then selects a subset of these partial solutions to pass on to the next iteration. This approach is competitive with NJ in recovering distinct low-cost topologies, while still being computationally efficient.

Materials and Methods

The NJ Method

The NJ method was initially proposed by Saitou and Nei (1987) and later modified by Studier and Kepler (1988). NJ seeks to build a tree that minimizes the sum of all edge lengths, i.e., it adopts the minimum-evolution (ME) criterion. (The ME criterion seeks the best phylogeny for the input distance matrix that minimizes the sum of all edge lengths, where edge lengths are assigned to minimized the least-squares deviation.) A number of studies have corroborated NJ's performance in reconstructing correct evolutionary trees (Saitou and Imanishi 1989; Kuhner and Felsenstein 1994; Huelsenbeck 1995).

Key words: phylogenetic reconstruction, neighbor-joining, least-squares, minimum evolution, solution space sampling.

Address for correspondence and reprints: William R. Pearson, Department of Biochemistry, Jordan Hall, Room 6-229, Charlottesville, Virginia 22908. E-mail: wrp@virginia.edu.

Mol. Biol. Evol. 16(6):806–816. 1999

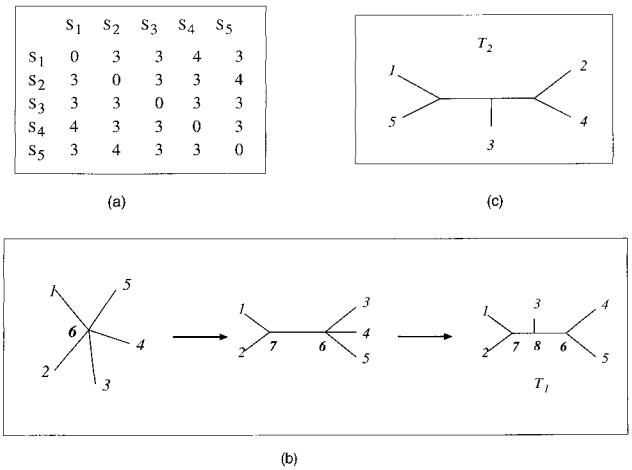

FIG. 1.—The NJ heuristic. *a*, An input matrix of size 5. *b*, An NJ solution strategy. First, the star tree topology centered at node 6 is formed; next, the closest neighbor pair {1; 2} is ''joined'' into a distinct internal node 7; finally, this new internal node 7, together with one of the leaf nodes 3, is joined to a new internal node 8 to form T_1. *c*, An equally good solution T_2, but with a very different topology, which was not found by NJ.

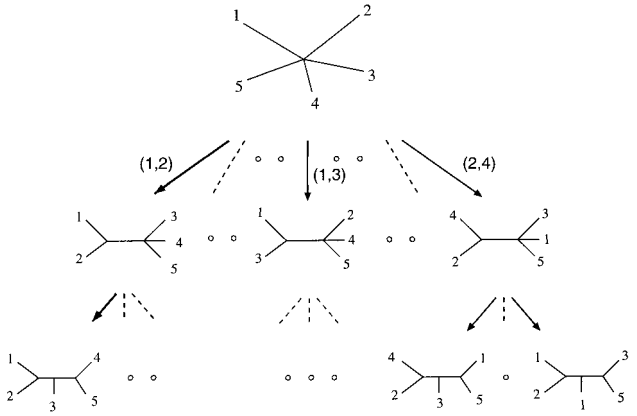

FIG. 2.—The GNJ method. The GNJ heuristic for the data of figure 1 *a* is shown. Throughout the search, three partial solutions are kept ($K = 3$). At each iteration, all possible neighboring taxa pairs are examined; three are selected to pass to the next iteration. The dashed lines represent the neighboring pairs that are eliminated during the current iteration. While NJ follows the path on the left and thus only finds the single tree T_1, the GNJ scheme recovers both equally good solutions T_1 and T_2 (see fig. 1).

For small numbers of taxa, NJ solutions are likely to be identical to the optimal ME tree (Saitou and Imanishi 1989).

NJ begins with a star tree, then iteratively finds the closest neighboring pair (i.e., the pair that induces a tree of the minimum sum of edge lengths) among all possible pairs of nodes (both internal and external). The closest pair is then clustered into a new internal node, and the distances of this node to the rest of the nodes in the tree are computed and used in later iterations. The algorithm terminates when $n-2$ internal nodes have been inserted into the tree (i.e., when the star tree is fully resolved into a binary tree). The NJ heuristic is illustrated in figure 1b.

Although the NJ method runs quickly, it returns only the single best solution found by its greedy search strategy. This solution can be further improved with postprocessing by rearranging branches and swapping subtrees (Rzhetsky and Nei 1992; Swofford 1996), but such improved solutions tend to remain topologically similar to the original starting-point solutions. To increase our confidence in the solution's reliability, it is natural to ask if there are other solutions, with different topologies, that are equally well supported by the distance matrix data.

Solution spaces can exhibit many alternate local optima (Penny et al. 1995). For instance, among all 15 possible trees of 5 taxa, 2 of them (T_1 in fig. 1b and T_2 in fig. 1c) fit the input matrix (fig. 1a) best. However, these two trees have very different topologies; they share no common internal edges. Indeed, according to the partition distance metric (see *Algorithms Compared*, below), T_1 and T_2 are the most dissimilar trees possible.

The Generalized NJ Method

Our generalized NJ (GNJ) method samples the solution space extensively by keeping track of multiple partial solutions as it progresses (the number of partial solutions K is an input parameter). Unlike the NJ method, which follows only a single path toward a solution, GNJ performs a more thorough search of the solution space by tracking and exploring many potentially good paths. That is, GNJ retains promising partial solutions, which may not be locally optimal but which have the potential for substantially greater cost savings in subsequent steps. An execution example of GNJ on the matrix of figure 1*a* is shown in figure 2.

The GNJ algorithm can select in several ways the K partial solutions that are passed on to the next iteration. A simple strategy would save the K best (i.e., least-cost) partial solutions; alternatively, partial solutions can be chosen at random. Selecting the best solutions tends to improve solution quality, while randomly selecting alternates tends to increase solution diversity. We implement a hybrid scheme that balances these two extremes: the top $Q \leq K$ least-cost solutions are selected, along with additional $D = K - Q$ ''topologically diverse'' solutions. (The parameter names Q and D are mnemonic for quality and diversity, respectively.) If there are more than Q least-cost pairs at a given iteration, GNJ will select Q of them arbitrarily, which makes the GNJ method nondeterministic.

To achieve topological diversity, at each iteration, after selecting the best Q partial solutions, the remaining partial solutions are partitioned into G groups according to their topological distances from a best partial solution (partial solutions within the same group are equidistant from the best partial solution). We then obtain an additional D ''topologically diverse'' partial solutions for the next iteration by selection the top $\lfloor D/G \rfloor$ solutions from each group. If G does not divide D, we select one additional solution from each of the $D - (\lfloor D/G \rfloor \cdot G)$ groups corresponding to the topological distances farthest from the best partial solution. Thus, at the last step toward solving a 16-taxon problem, alternate solutions can be as many as 13 partitions away from the best

current solution. In this case, if $D = 50$, at least $\lfloor 50/13 \rfloor = 3$ best solutions at topological distances 1 and 2 are saved, and at least the 4 best solutions at topological distances 3 through 13 are saved. For 32 taxa and $D = 100$, at least 3 solutions will be saved at each topological distance. Because the maximum topological distance increases linearly with the number of taxa, the above strategy ensures that the number of topologically distinct good solutions can remain relatively constant by increasing D linearly, rather than exponentially, with the number of taxa.

A similar idea is employed in the stepwise ME tree-building method (Kumar 1996). At each iteration, for a given partial tree, this method first identifies the leading node (i.e., the node most likely to be joined to another node), and forms the set of next-step NJ trees by clustering each node with the leading node. This strategy restricts the solution space somewhat, but it requires exponential time to run, which makes it practical only for small data sets. Moreover, it does not explicitly consider alternate solutions at different topological distances (see below), so it is less likely to identify topologically distinct alternatives.

Different combinations of Q and D ($K = Q + D$) enable a smooth trade-off between quality versus diversity. As Q increases with respect to D (for a fixed K), lower-cost solutions are favored over ones with diverse topologies, while for smaller values of Q, the solution space exploration becomes broader, and topologically different local optima are more likely to emerge. We note that if $K = Q = 1$ (and thus $D = K - Q = 0$), the single solution returned by our GNJ approach is identical to the solution produced by the original NJ method (Saitou and Nei 1987; Studier and Keppler 1988). Here, only the best-cost partial solution is passed to each subsequent iteration, which is exactly what NJ does. Thus, GNJ directly generalizes the NJ method.

Additional strategies for expanding the search of phylogenetic tree space might be considered. The GNJ approach can be abstractly divided into two phases: (1) a tree generation component which produces multiple partial solutions, and (2) a partial solution evaluation function which favors certain preferred partial solutions over others. The overall run time per iteration of the combined method is asymptotically no greater than the slowest of these two components.

The algorithm described in this section utilizes the NJ method as the partial tree generation mechanism in phase 1, while using the ME criterion (implicit in the NJ method) in filtering candidate partial solutions in phase 2. However, any combination of existing algorithms or heuristics for tree generation and tree evaluation can be incorporated into this general template. For example, we can evaluate partial trees at each step using the least-squares deviation optimality criterion. (The least-squares criterion seeks the best phylogeny for the input distance matrix that minimizes $\Sigma_{1 \le i \le j \le n} (t_{ij} - d_{ij})^2$, where d_{ij} is the distance between taxa i and j in the input distance matrix, and t_{ij} is the sum of all the branch lengths along the unique path connecting taxa i and j in the postulated phylogeny.)

An alternative scheme for tree generation might allow arbitrary partitions at intermediate steps (i.e., ''join'' any number of taxa rather than exactly two). In this case, a number of existing efficient partitioning heuristics (Alpert and Kahng 1995) can be readily applied to generate more promising and diverse partial solutions. Likewise, the method for selecting topologically diverse partial solutions might select more solutions from more distant topologies, rather than uniformly sampling the topological distances as is done in this implementation.

The GNJ program is written in the C programming language and is available from ftp://ftp.virginia.edu/pub/fasta/GNJ. To make the GNJ results more usable in practice, we output the trees obtained by GNJ in a computer-readable format that can be readily processed by other programs (e.g., the `consense` program in the PHYLIP package). Moreover, we summarize the leaf partitions found among the GNJ solutions below a threshold cost and rank them by decreasing frequencies.

Data Sets

We tested the GNJ heuristic in the UNIX environment. Two types of distance matrices were used to evaluate the algorithm:

(1) Distance matrices were constructed for nucleotide sequences generated by randomly mutating an ''ancestral'' sequence along a model evolutionary tree using the treeDNA program (J. Felsenstein, personal communication) with the Kimura (1980) two-parameter model for mutation rates. Three types of topologies were used for the model trees: topologies of minimum diameter (which we refer to as type A), topologies of maximum diameter (type B), and a mixture of both (type C). Here, the diameter of a topology is defined as the maximum number of edges connecting any two leaf nodes within the topology. Therefore, topologies of type A are the most ''branchy'' (i.e., they resemble a complete binary tree), while topologies of type B are more ''stringy.'' Type A trees were the most challenging and are used for most of the figures.

Divergence rates ranging from 0.005 (internal branches) to 0.50 (leaf or external branches) were used to produce the synthetic data. Two different type A and type B data sets were examined. Type A1 and type B1 data sets used divergence rates $\sim$0.02 (32 taxa) to $\sim$0.05 (8 taxa) for internal edges and 0.4 for external edges (thus, the ratio of external to internal branch rates varied from 10 [for 8 taxa] to 35 [for 32 taxa]). Type A2 and type B2 trees used rates of 0.005 for the central (internal) edges and 0.50 for the external (leaf) edges (external/internal ratios of 100).

(2) Several biological data sets were examined, including immunological data from 9 frog species (Saitou and Nei 1987), data from 13 viral *env* V3 fragments and *gag* P17 (Leitner et al. 1996), and 47 aligned TCP-1 chaperonin 60 family members (J. S. Blandfort, personal communication). For DNA sequences, the distance matrices were computed with the `dnadist` program in the PHYLIP package (Felsenstein 1993), using the Kimura (1980) two-parameter model. For protein sequences, the

distance matrices were computed with the `protdist` program in the PHYLIP package (Felsenstein 1993), using the Dayhoff (1978) PAM matrix model. We obtain 30 biological data sets of 8, 16, or 32 taxa by randomly sampling the original data sets. Results on the different biological data sets were similar; only results on the chaperonin distances (referred to as data set R1) are reported.

Algorithms Compared

We evaluated the data sets using three algorithms: (1) the NJ method (Saitou and Nei 1987; Studier and Keppler 1988), as implemented in the PHYLIP package (Felsenstein 1993); (2) the Fitch-Margoliash (FM) method for fitting topologies to distance matrices with respect to the least-squares criterion (Fitch and Margoliash 1967), as implemented in the PHYLIP package (Felsenstein 1993); and (3) the GNJ method, described in this paper.

In addition, we examined every possible tree topology for synthetic and biological data over eight taxa. This exhaustive method is guaranteed to return a global optimum (i.e., the lowest-cost topology). Because of the sheer size of the solution space, the optimal method is feasible only for data sets containing fewer than 10 taxa.

The solutions from the different algorithms were evaluated using either the least-squares or the ME criterion. Least-squares tree cost is computed by assigning nonnegative edge lengths in a way that minimizes the least-squares deviation. ME tree cost is computed as the sum of such edge lengths in a tree.

To further improve the solution quality, we also applied a postprocessing optimization step which rearranges subtrees as follows. Given a topology, we compute the cost of all the trees resulting from swapping/exchanging subtrees around each of the internal edges of the topology. Then, the lowest-cost tree is chosen as the new current tree, and its neighborhood is investigated in turn. We iterate this process until no further improvement can be obtained.

Topological distances in this paper are based on the partition metric (Robinson and Foulds 1981; Penny and Hendy 1985; Steel and Hendy 1993), which measures the number of edges common to a given pair of binary trees. Each internal edge naturally partitions the set of leaf nodes into two subsets. Two trees spanning the same set of leaves have a common edge if removing this edge induces the same two subsets of leaf nodes. Thus, the partition distance between any two trees is defined as the number of edges in one tree for which there are no corresponding equivalent edges in the other tree. Since each binary tree of n leaves has $n - 3$ internal edges, distances under the partition metric can be represented as integers between 0 and $n - 3$.

Results

Like NJ, GNJ seeks to identify phylogenetic tree topologies and branch lengths that best fit distance data. GNJ improves on NJ by identifying near-optimal topologies that are significantly different from the best solution found in the search (there are typically many near-optimal solutions that differ only slightly from the best solution; we seek topologically distant alternatives). In the results below, we first show that the data sets that we examine contain topologically distinct low-cost solutions. We then demonstrate that the GNJ algorithm can find these low-cost alternative solutions; by examining two measures of success: (1) the number of alternative trees found by GNJ with a near-optimal cost and (2) the maximal topological (partition) distance between the near-optimal solutions and the optimal solution found. (In the case of more than eight taxa, where an exhaustive search for the optimal solution is computationally infeasible, we compare with the best solution found instead of with the optimal solution.) In both tests, we seek the largest number of solutions with cost nearest to optimal but with topological distance that is far away.

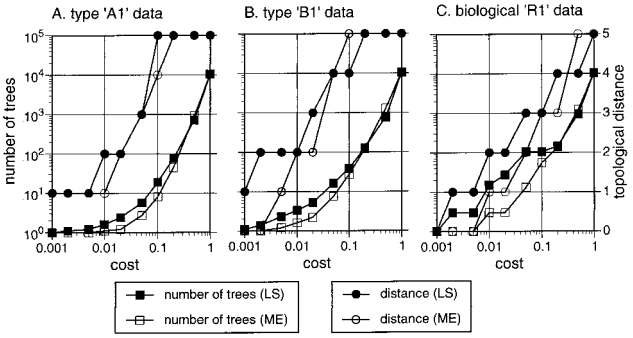

FIG. 3.—Distribution of tree costs and diversity. The number of trees (left ordinate, ■, □) and maximum topological (partition) distance averaged over 30 data sets (right ordinate, ●, ○) are plotted as a function of the fractional cost range. Distributions for the least-squares (LS, ■, ●) and the minimum-evolution (ME, □, ○) criteria are shown. Distributions were determined for eight taxa from 30 synthetic type A1 data sets, 30 synthetic type B1 data sets, and 30 data sets from biological data set R1. The figures show the results determined after an exhaustive search of all 10,395 tree topologies for 8 taxa.

Comparison of GNJ with Exhaustive Eight-Taxon Searches

To judge how effectively the GNJ approach finds alternative topologically distinct solutions, we first characterized the actual number and diversity of near-optimal solutions by enumerating all 10,395 different trees for data sets with eight taxa and calculated the cost for each tree topology (fig. 3). Tree costs were optimized using either the ME criterion or the least-squares criterion. Because the different cost criteria may have different distributions of costs, we plot the number of trees obtained as a function of the fractional cost range: $(c_x - c_{\min})/(c_{\max} - c_{\min})$, where c_x is the least-squares or ME cost of a specific tree topology, $c_{\min}$ is the minimum (and for exhaustive searches, optimal) cost under that criterion, and $c_{\max}$ is the cost of the worst topology. For the eight-taxon data, $c_{\min}$ and $c_{\max}$ are known because the cost of every possible topology has been calculated. For larger data sets, $c_{\min}$ is approximated from the minimum cost obtained for all of the tree searches on the data set, and $c_{\max}$ is approximated from the maximum cost obtained by sam-

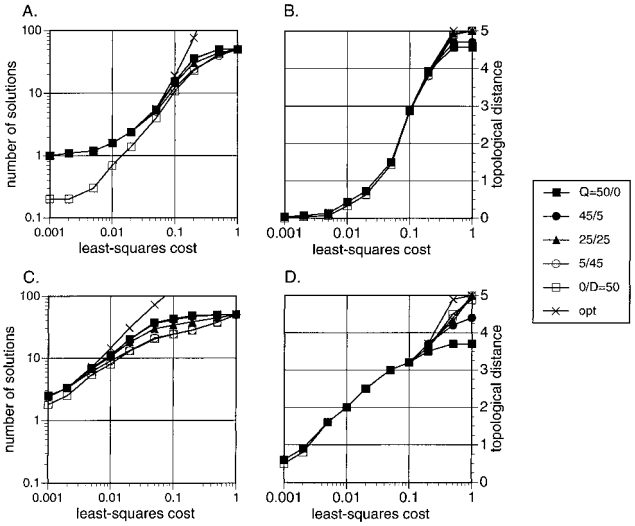

FIG. 4.—GNJ solutions—eight taxa. The distribution of solutions found by GNJ on 30 type A1 synthetic eight-taxon data sets (*A* and *B*) or 30 R1 biological eight-taxon data sets (*C* and *D*) are shown. Searches were done with $K = Q + D = 50$. Panels *A* and *C* show the average numbers of different trees with costs within the fractional least-squares cost shown. Panels *B* and *D* show the averages of the maximum topological distances of the solutions within the fractional cost range.

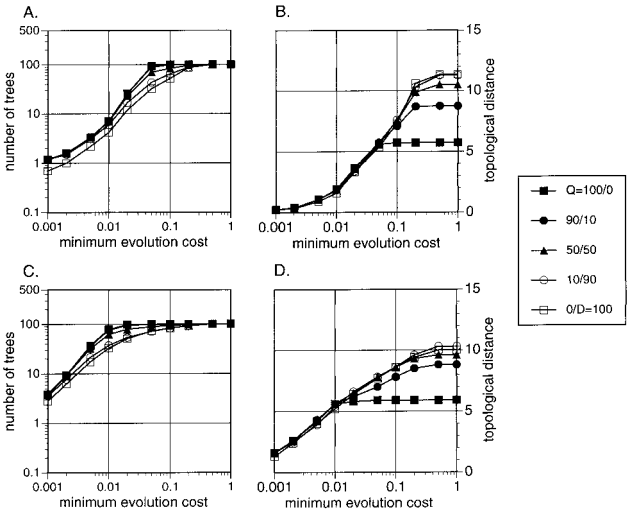

FIG. 5.—GNJ solutions—16 taxa. The distribution of solutions found by GNJ on 30 type A1 synthetic 16-taxon data sets (*A* and *B*) or 30 R1 biological 16-taxon data sets (*C* and *D*) are shown. Searches were done with $K = Q + D = 100$. Panels *A* and *C* show the average numbers of different trees with costs within the fractional ME cost shown. Panels *B* and *D* show the averages of the maximum topological distances of the solutions within the fractional cost range.

pling 100 trees randomly. Thus, for the 16- and 32-taxon data sets, c_{min} may not be the optimal minimum cost and c_{max} may not be the highest (worst) cost, but these approximations should differ only slightly from the true values.

For the synthetic data, it is possible to ask how often the low-cost trees found by the GNJ algorithm were consistent with the original tree that was used to produce the distance data. However, the lowest-cost least-squares or ME tree was often different from the original tree. Trees from type A1 and B1 data are used for most of the figures because the difference between the original tree cost and the best tree cost was typically between 0 and 0.1, with the median between 0.01 and 0.03 of the cost range. Trees from type A1 and B1 synthetic data behaved very similarly to trees from the biological data sets. For the A2 and B2 data sets, the median original tree cost was 0.4–0.7 of the cost range. Thus, because of the high external/internal rate ratio, the best tree frequently had a cost substantially lower than that of the original tree, and these data sets have a large number of distinct local minima, which are not seen with the biological data sets or with the type A1 and B1 trees.

Figure 3 shows how the number of trees and the topological distance between the alternative solutions increases over the fractional cost range. The results from three different data sets are shown using either the ME or the least-squares cost criterion. In these plots, more challenging data sets have a larger number of near-optimal trees and greater topological distance at lower fractional cost. In general, there are more near-optimal trees with the least-squares criterion than with the ME criterion, and those trees tend to be more topologically distinct (fig. 3). For example, with the biological data (fig. 3*C*), there were 14.6 trees on average with least-squares cost within 0.01 of optimal, but only 2.6 trees ≤ 0.02 when the ME cost is calculated. Furthermore, when the cost is less than 0.01, the maximum topological distance for near-optimal trees is greater for the least-squares trees than for the ME trees.

The branchy type A1 synthetic data set tends to produce a larger number of near-optimal, topologically distant trees than the type B1 (fig. 3) data sets. When the type A2, B2, and C2 data sets were examined (data not shown), type A2 data sets were the most challenging, and, for trees with costs of ≤ 0.01, number of trees and topological distance between the trees were about twice as high for type A2 as for type A1. The biological data set appears more challenging than the type A1, B1, and B2 synthetic data sets, but less challenging than the type A2 data set (fig. 3 and data not shown). We focus our attention on the number and diversity of trees with cost range 0.01–0.05 both because these cost ranges are intuitively close—between 1% and 5% of the best cost found—and because, for the type A1 and B1 synthetic data, 0.01–0.05 spans the range of cost differences between the original trees used to generate the distance data and the best trees found for the data.

Ideally, the GNJ algorithm would find each of the near-optimal solutions that can be found when every tree topology is examined. Thus, we use the number of solutions, their average cost, and their diversity to gauge the effectiveness of GNJ (figs. 4–6) and compare GNJ with an exhaustive search (fig. 3). We seek combinations of Q and D that approach the distribution of solutions seen in the exhaustive search. Figure 4*A* shows that the GNJ algorithm effectively identifies virtually all suboptimal solutions with costs of ≤ 0.05 on the synthetic data set as long as $Q > 0$. (Results, not shown, using the ME cost criterion are indistinguishable.) Only when

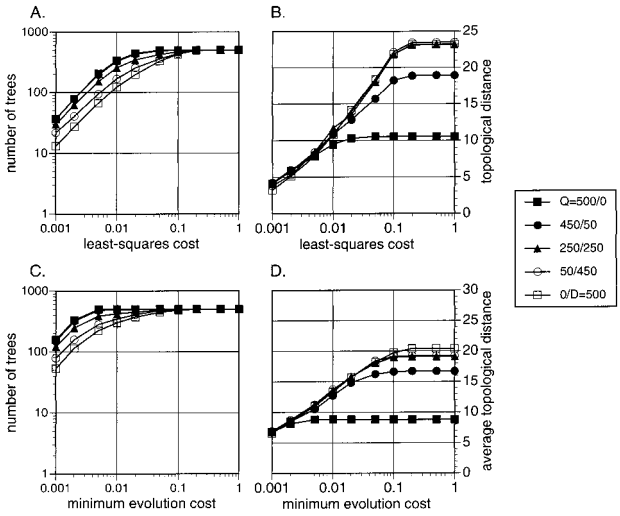

FIG. 6.—GNJ solutions—32 taxa. The distributions of the number of trees (*A*) and the maximum topological distance averaged over 30 data sets (*B*) are shown for the least-squares cost criterion on 32 taxa synthetic type A1 data. Searches were done with $K = Q + D = 500$. Panels *C* and *D* show tree numbers and topological distances for the biological R1 data.

$Q = 0$ and $D = 50$ is the number of near-optimal solutions different from the number found in the exhaustive search; that is, some of the lowest-cost alternative solutions are missed. The curves in figures 4*B* and *D* report the average maximum topological distance; i.e., the maximum topological distance among all the trees with cost less than the ordinate is determined for each of the 30 data sets, and the 30 maximum distances are averaged. Again, when $Q > 0$, the alternate solutions found by the GNJ algorithm are as diverse as those found by the exhaustive search for costs within 10% of optimal. (We also examined the maximum topological distances for the data in fig. 4 and found that they were very similar to those for the exhaustive search if $Q > 0$; data not shown.)

The biological R1 data set is more challenging in some ways—there are a larger number of alternate solutions with low cost (fig. 4*C*), and the low-cost solutions appear to be more topologically diverse (fig. 4*D*). For the biological data set, GNJ begins to miss solutions with costs of >0.005 that are found by the exhaustive search. At a fractional cost of 0.01, 11 of 15 solutions are found by GNJ with $Q \geq 25$, and 18 of 31 are found at fractional cost ≤0.02. As with the synthetic data set (fig. 4*B*), when $Q = 0$, some of the best near-optimal solutions are missed. The results presented in figure 4 suggest that for small (eight-taxa) problems, the GNJ algorithm identifies alternate near-optimal, topologically distant solutions very effectively.

GNJ Performance with 16 and 32 Taxa

For larger data sets, it is not computationally feasible to examine the solution space exhaustively, so we cannot directly compare the GNJ results with the optimal solution. (Likewise, we cannot guarantee that the lowest-cost solution is optimal, but it is likely to be near optimal.) Nonetheless, we can still evaluate how the GNJ algorithm benefits from saving multiple $K = Q + D$ solutions by examining a range of Q, D pairs (figs. 5 and 6). When type A1 synthetic data sets with 16 taxa are searched, the largest number of low-cost solutions are again found when $Q > 0$, and the most topologically diverse solutions are found when $D > 0$. (Fig. 5 shows the results using the ME cost criterion; results using the least-squares criterion, not shown, are similar.) For these data sets with $K = 100$, the trade-off between quality Q and diversity D is clear-cut. Below 0.01, there is little difference in diversity as Q and D change; above 0.02, $D \geq 50$ gives the best results. On the biological data set (fig. 5*C*), searches with $Q \geq 100$ find almost twice as many (62–78) solutions with fractional cost ≤0.01 as do searches with $D = 90$ or 100 (33–37 solutions). The difference in performance with respect to Q and D increases at higher fractional costs. However, while reducing D increases the number of low-cost solutions found, it also decreases the diversity of the solution set. For these data, $Q = D = 50$ seems to be the best compromise.

When searches are performed on 32-taxon data (fig. 6), the importance of D in improving the diversity of the solutions is more apparent. As before, solutions with $Q = D = 250$ appear to provide a good balance between finding the largest number of low-cost solutions and finding the most diverse solutions. We note that as Q increases from 250 to 500, there is little change in the number of trees with fractional cost ≤0.02 on the synthetic type A1 data set (fig. 6*A*), and that the maximum topological distance among those solutions increases very little as D increases from 250 to 500. Thus, for these synthetic data, although $K = 500$ retains only a tiny fraction of up to 10^{40} possible 32-taxon tree topologies, the data in figures 6*A*–10 suggest that most of the lowest-cost solutions, and many of the topologically diverse solutions, are found.

Comparison with Other Methods

Thus far, our results suggest that GNJ can identify alternative near-optimal solutions when K ranges from 50 (8 taxa) to 200 (32 taxa). In this section, we compare GNJ with different $K = Q + D$ values with two popular phylogenetic tree reconstruction methods for distance data, the NJ method (Saitou and Nei 1987) and the FM algorithm (Fitch and Margoliash 1967) as implemented in the PHYLIP package (Felsenstein 1993). As before, we consider both synthetic and biological data sets with different numbers of taxa, and we compare two cost criteria: the ME criterion used for NJ searches, and the least-squares criterion used by FM. In these tests, we again consider two measures of success: quality (cost) and diversity. We evaluate the quality of the solutions in two ways: (1) by the fraction of the time (for the 30 test data sets) that a near-optimal solution is found and (2) by the average cost of the best solutions found. To evaluate diversity, for each distance matrix, we first compute the maximum topological distance between pairs of near-optimal GNJ solutions. Diversity is then measured by computing (1) the maximum, as well as (2) the average of these distances, over 30 data sets.

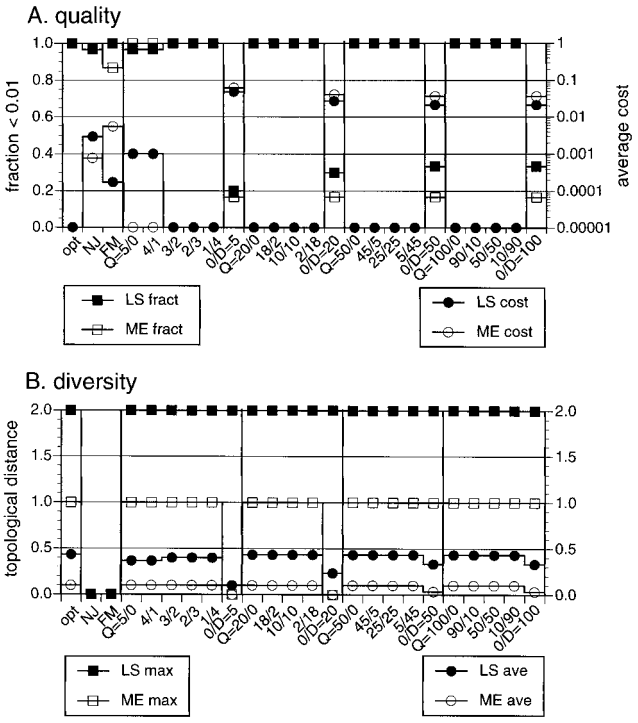

FIG. 7.—GNJ performance—eight taxa. The quality and diversity of GNJ solutions with different values of Q and D are compared with the optimal solution set (opt) and with NJ and FM solutions for synthetic type A1 data *A*, The fraction of the time a solution was found with a cost < 0.01 of optimal (squares, left axis) using either the least-squares (filled symbols) or the ME (open symbols) criterion. The right axis (circles) reports the cost of the best solution found, averaged over the 30 data sets. *B*, The diversity of the $K = Q + D$ solutions with costs of <0.01 is shown as the largest topological diversity found among all 30 data sets (squares) and the maximum topological diversity averaged over the 30 data sets. Closed symbols report diversity for least-squares solutions; open symbols report ME diversity.

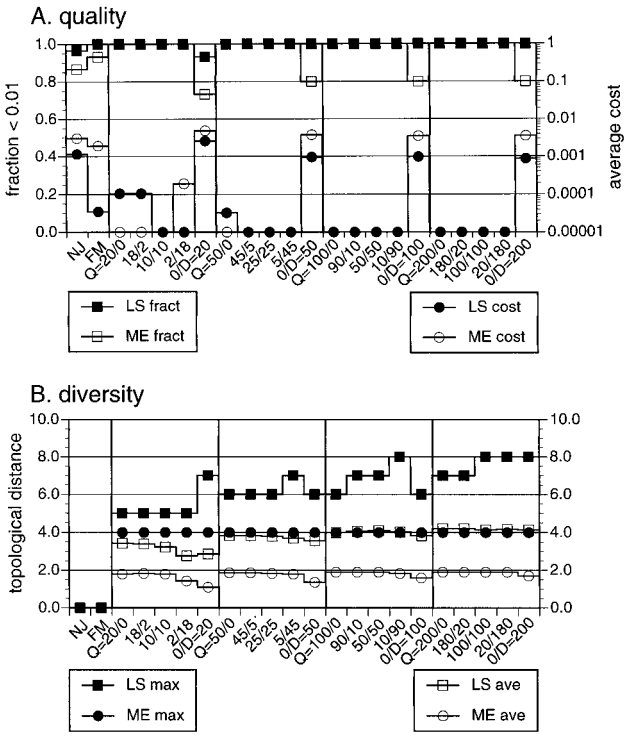

FIG. 8.—GNJ performance—16 taxa. Results for 16-taxon type A1 synthetic data are plotted as in figure 7, except that both the fraction of solutions found and the topological distance plot use a cost threshold of 0.01.

For eight-taxon type A1 data, GNJ finds solutions of very high quality that are as diverse as the exhaustive search when $K > 5$ and $Q > 2$ (fig. 7). When $Q > 2$, a solution within a cost range of 0.01 of optimal is found 100% of the time. For $Q = 0$, GNJ finds a solution within 0.01 of optimal less than 20% of the time when $D = 5$, and less than 40% of the time when $D > 5$. On the same data sets, NJ finds a <0.01 ME solution and FM finds a <0.01 least-squares solution more than 95% of the time. The average cost data in figure 7*A* show that the best solutions found by NJ and FM are typically within 0.01 of the cost range, but those found by GNJ ($K \geq 20$ and $Q \geq 2$) are optimal. Thus, GNJ consistently finds solutions with costs lower than both NJ and FM. Moreover, comparison of both the largest maximum topological distance and the average maximum topological distance (fig. 7*B*) shows that when the optimal solution was found by GNJ, the diversity of solutions found (with costs <0.01 of optimal) is as large for the GNJ solution set as for those found by the exhaustive search. GNJ performed as well as the exhaustive search on the much more challenging type A2 data as well (not shown).

Results for 16 taxa are shown in figures 8 and 9. On the synthetic type A1 data set, GNJ found a solution

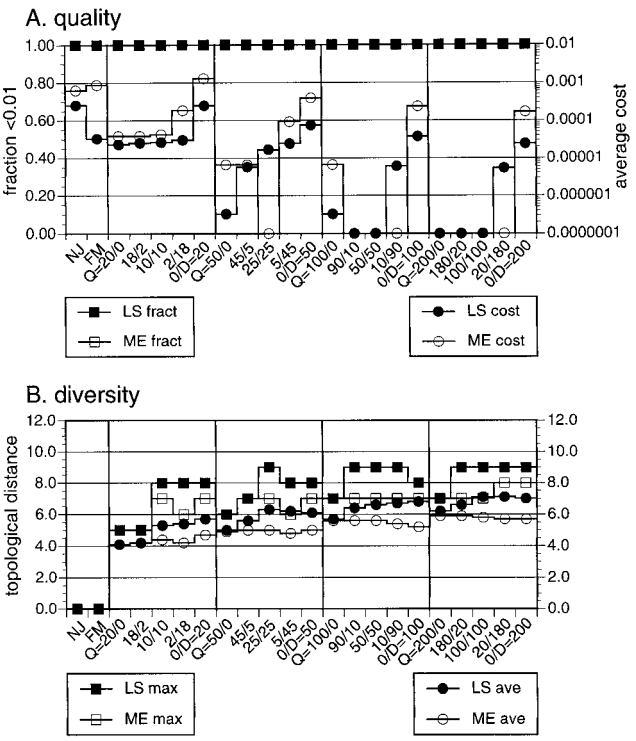

FIG. 9.—GNJ performance—biological data. Results for 16 taxa from biological data set R1 are plotted as in figure 8.

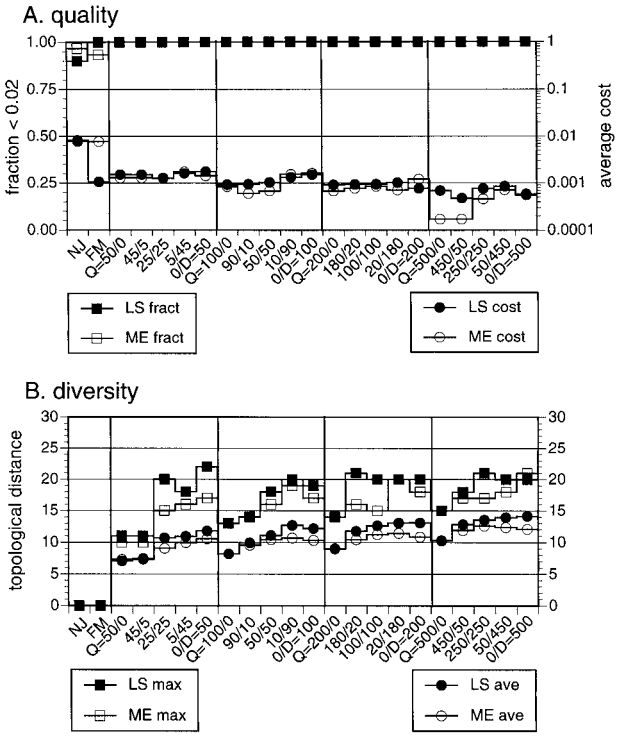

FIG. 10.—GNJ performance—32 taxa. Results for 32-taxon type A1 synthetic data are plotted as in figure 8, except that both the fraction of solutions found and the topological distance plot use a cost threshold of 0.02.

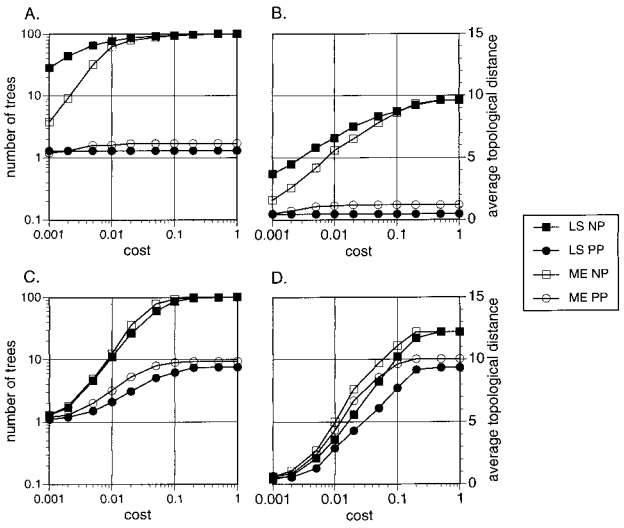

FIG. 11.—Number and diversity of postprocessed solutions. Results for 16-taxon biological R1 data (*A* and *B*) and 16-taxon type A2 data (*C* and *D*) are plotted as in figure 5. Results shown are for $K = 100$, $Q = D = 50$, with either nonpostprocessed (squares) or postprocessed (circles) solutions. Filled symbols report the distribution of least-squares solutions with fractional cost; open symbols plot ME costs.

within 0.01 of the best cost 100% of the time when $Q > 0$. For these data, NJ found a <0.01 cost solution only 80% of the time using the ME criterion, while FM always found a <0.01 cost solution. Once again, GNJ found solutions with lower average costs. For type A2 data (not shown), NJ and FM found <0.01 solutions only 30%–55% of the time for the least-squares criterion, and 25%–37% of the time for the ME criterion, while GNJ found the best ME solution 100% of the time when $Q > 5$. GNJ found the best least-squares solution more than 80% of the time on type A2 data with $Q \geq 5$. As K increased from 20 to 200, the cost of the best solutions consistently improved with GNJ. While we cannot compare the GNJ diversity with the diversity that would be found by an exhaustive search, increasing K from 20 to 200 improves the average maximum diversity, and, as before, $Q = D$ seems to provide low-cost solutions with high diversity.

When 16-taxon biological data are examined, the NJ and FM algorithms perform quite well (fig. 9). However, even with these data, the average cost of the best solutions found improves from about 10^{-4} to 10^{-6} when GNJ is used and $Q \geq 25$.

NJ, FM, and GNJ all perform well on 32-taxon type A1 (fig. 10) and biological data (not shown) using a cost threshold of 0.02 or 0.05 (not shown). However, it is surprising how diverse the GNJ solutions are; when solutions with costs within 2% of the best cost are included, GNJ found alternate low-cost solutions that share fewer than half of the internal edges (two trees share an internal edge if the edge induces the same leaf bipartitions in both trees).

Comparison of the cost and the diversity of GNJ solutions with $K = 200$ and $K = 500$ suggests that $K = 500$, which increases the run time 2.5-fold, is probably unnecessary, since neither the quality of the solutions nor the diversity increases significantly with the higher K. Again, using $Q = D$ provides a good balance of quality and diversity.

Postprocessing

Rzhetsky and Nei (1992) have observed that for small data sets, NJ solutions are likely to be topologically close to the optimal solution. We examined how postprocessing (described in *Algorithms Compared*) affects the number and diversity of the low-cost solutions, and how postprocessing might improve NJ, FM, and GNJ-based initial solutions. The postprocessing algorithm examines all the trees that can be formed by swapping (exchanging) subtrees around each of the internal edges in the tree, thus considering all the alternative trees that are within one partition distance from the initial tree. If a topology is found with a lower cost (least-squares or ME), the process is repeated, until no topological neighbor is found with a lower cost. If the GNJ algorithm finds alternate solutions that are on different sides of a single shallow cost basin, postprocessing should reduce the number and diversity of low-cost alternate trees. This seems to be the case for the biological R1 data (fig. 11*A*) and the synthetic type A1 data (similar to the biological R1 data; not shown). Alternatively, if GNJ actually finds distinct local minima (with respect to cost), the number of trees may decrease dramatically, but the topological distance between alternate solutions should remain substantial. Multiple distinct local minima are found with the synthetic type A2 data.

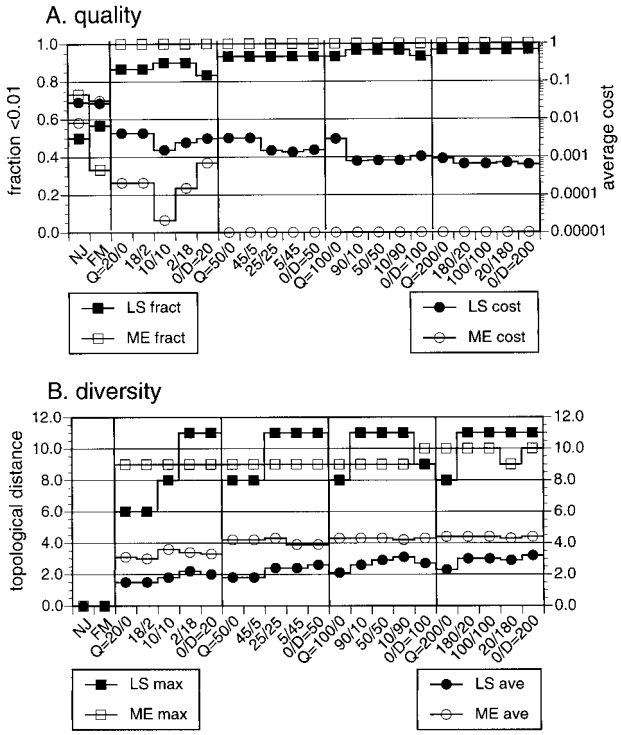

FIG. 12.—Postprocessed NJFM, and GNJ performance. Postprocessed results for 16-taxon type A2 data are plotted as in figure 8.

The results of postprocessing on the 16-taxon data sets suggest that GNJ is capable of identifying alternate topologically distinct local minima when they exist (fig. 11). As expected, the number of distinct solutions drops dramatically (because of convergence) when the GNJ solutions are postprocessed. For the biological R1 data (fig. 11*A* and *C*), the drop is more than 30-fold, as it is with the synthetic type A1 data (not shown). However, for the synthetic type A2 data, which is derived from trees in which the cost for the original tree is frequently midway between the best and worst costs, the drop is only two to three fold, and the average maximum topological distance drops only about 20%. Thus, for this very difficult data set, many of the alternate solutions found by GNJ cannot be reached by local branch-swapping from the best solution, and distinct local minima have been found. Figure 12 compares the performances of NJ, FM, and GNJ, each followed by postprocessing, on 16-taxon type A2 data. Postprocessing improves the performance of NJ and FM in finding a solution with cost <0.01 from about 30%–50% success to 50%–70% success; GNJ is 100% successful with every combination of *K, Q,* and *D.* Again, GNJ finds lower-cost solutions that NJ and FM fail to find even after exhaustive postprocessing. For these data, NJ and FM appear to sometimes find local minima (with respect to cost), while GNJ finds more global minima.

The average maximum topological diversity for the difficult type A2 data decreases only slightly with postprocessing, and the maximum topological diversity is as high after postprocessing as before. This result—topologically diverse solutions despite a dramatic decrease in the number of low-cost solutions—implies that GNJ has found alternate local minima that cannot be reached by local branch swapping from the lowest-cost solution. Since our postprocessing strategy begins from the *K* solutions found by conventional GNJ, GNJ without postprocessing can detect topologically distinct alternative local minima. For the synthetic type A2 data sets, low-cost solutions that are topologically distinct local minima appear often. For example, for $Q = 50$ and $D = 50$, the least-squares solutions differing by 11 (out of a maximum 13 possible) branch swaps (partitions) are found in at least one data set, and they differ by 2.9 swaps on average (ME solutions differ by as much as 9 partitions, and by 4.3 on average). This suggests that topologically distinct solutions have been found by GNJ on 25%–50% of the 30 synthetic data sets.

Table 1
Execution Times

Method	8 taxa	16 taxa	32 taxa
NJ	<0.01	0.01	0.09
FM	0.08	2.4	31.5
GNJ			
$K = 20$	0.08	0.8	9.8
$K = 50$	0.2	2.1	25.1
$K = 100$	0.5	4.4	52.1
$K = 200$	1.1	8.8	103.7
$K = 500$	3.1	24.2	262.7

NOTE.—Run time (in CPU seconds on a 167-MHz UltraSparc) for Neighbor-Joining (NJ), Fitch-Margoliash (FM), and GNJ, averaged over 100 type A input data sets of various sizes, using the least-squares criterion and $Q = D$.

The results with the biological and synthetic type A1 data sets contrast starkly to the diversity found with the synthetic type A2 data. With the least-squares criterion and postprocessing, the average maximum topological diversity is about 0.5 and the maximum diversity is 4, implying that distinct solutions are found in only about 10% of the data sets. With the ME criterion, the maximum diversity is the same, but the average maximum is 1.1; again, topologically diverse solutions may be found for 25% of these 16-taxon data. For the synthetic type A1 data, the average diversity drops from about 4 to 1.1 (least-squares) or from 7 to 2 (ME).

Although postprocessing can improve the quality of GNJ solutions without significantly reducing their diversity, the time required to postprocess *K* alternative solutions can be prohibitive when the number of taxa (and thus the number of branch swaps that must be tested) is large (>16). However, comparison of fig. 12*A* and the nonpostprocessed data (not shown) suggests that postprocessing does not improve the solution quality significantly when Q and $D \geq 50$, and thus the extra computation is unnecessary.

Run Time

GNJ uses computation time roughly proportional to the number of partial solutions maintained during execution (*K*) and cubic in the number of taxa analyzed. Average run times of NJ, FM, and GNJ for various input sizes are shown in table 1. GNJ is considerably slower than NJ (which is one of the fastest tree construction algorithms available, because it does not evaluate any

alternative trees), and three- ($K = 200$) to eightfold ($K = 500$) slower than FM for the 32-taxon data sets.

During its execution, GNJ keeps track of K partial solutions. At each iteration, as the next pair of taxa is removed from the ''star'' tree, GNJ explores all the candidate solutions derivable from the current K partial solutions via a single NJ step. Since each partial tree induces $O(n^2)$ candidate trees by grouping one of the $O(n^2)$ possible node pairs in the tree, the cost of all the resulting candidate trees requires $O(n^2)$ evaluation time. Therefore, each GNJ iteration requires $O(Kn^2)$ time to examine the cost of all $O(Kn^2)$ candidate trees.

At each iteration, GNJ must also select K candidate trees to pass on to the next iteration. In this version, the selection process requires all $O(Kn^2)$ candidate trees to be sorted by cost. Currently, the time required by each iteration of GNJ is dominated by the sorting time, which is $O(Kn^2 \log (Kn^2)) = O(Kn^2 (\log K + \log n))$. We anticipate that the amount of data to be sorted can be reduced and that in future versions, the GNJ cost calculation will dominate the run time. Since GNJ has a total of $n - 3$ iterations, the overall run time for GNJ is $O(Kn^3 (\log K + \log n))$.

Discussion

The GNJ algorithm is explicitly designed to explore broadly phylogenetic tree solution spaces and seek low-cost solutions that are topologically distant. To achieve this goal, GNJ maintains multiple partial solutions at each iteration and incorporates both quality (tree cost) and diversity (topological distance) in selecting the set of partial solutions that will be passed on to the next iteration. The solution space sampling is controlled by the parameters K, Q, and D, which specify the number of partial solutions to retain and the balance between quality (Q) and diversity (D) in selecting alternate solutions.

For small data sets (e.g., 10 taxa), GNJ can perform better than NJ and FM algorithms by maintaining $K =$ 20–50 partial solutions. For example, for biological data sets over nine taxa, all eight trees of cost close to the NJ cost (under the least-squares criterion) are obtained by GNJ while maintaining only $K = 50$ partial solutions. For synthetic data sets of eight taxa, GNJ finds the optimal solution whenever $K \geq 20$ and $Q \geq 2$.

For data sets with 16 or 32 taxa, both the topological diversity and the quality of GNJ solutions improves as K increases. For the data sets that we examined, low-cost solutions were efficiently found with $K \geq 20$ for 8 taxa and with $K \geq 50$ for 16 taxa. Increasing K did little to improve either the quality or the diversity of the solutions. Thirty-two-taxon problems are far more challenging (and more common in the molecular biology literature). While $K \geq 200$ (32 taxa) was effective in finding low-cost solutions, increasing K improved the quality and, to a lesser extent, the diversity of the 32-taxon solutions.

We believe that the postprocessing results show that GNJ is capable of identifying low-cost, topologically distinct solutions that cannot be found simply by successively examining every topology near to individual low-cost trees. ''Falling into local minima'' is an inherent flaw of any phylogenetic search method that examines only a small portion of the solution space. The postprocessing results for the biological R1 data suggest that this data probably has a single, very broad local minimum with many different low-cost topologies but very few, if any, alternate solutions that cannot be found by postprocessing (local branch swapping). In contrast, the synthetic type A2 data do appear to have several distinct local minima, which were found by GNJ. While it is reassuring to learn that GNJ is capable of finding alternative local minima when they exist, more extensive simulations will be required to characterize the conditions under which large numbers of distinct local minima occur. While postprocessing may not be necessary to find high-quality solutions, the decrease in diversity with postprocessing should improve our confidence that a data set does not have many topologically distinct low-cost solutions.

Our results suggest that GNJ performs best when $Q = D = K/2$, and that $K = 200$ provides an excellent balance between computation time and solution quality/diversity for up to 32 taxa. For more than 50 taxa, $K =$ 500 or 1,000 may provide better solutions; however, this will depend greatly on the structure of the phylogenetic tree solution space. For large numbers of taxa, one can judge whether a larger K is likely to provide novel solutions by performing searches with $K = 100$ and 200. If $K = 200$ does not find any low-cost trees that were missed with $K = 100$, it is unlikely that $K = 500$ (or more) will uncover additional novel trees either.

GNJ is considerably slower than traditional NJ and for large problems (>32 taxa and $K \geq 200$), it is much slower than FM as well. However, a GNJ run is more accurately compared with multiple FM searches where the taxa are successively added in different orders, a process that can easily increase the amount of time required 20- to 50-fold. Because GNJ explicitly seeks out topologically diverse solutions, we believe that it is more likely to identify distinct alternatives than additional FM trials.

This paper considers the generalization of the NJ partitioning strategy to which the distance cost measures seem ideally suited. However, the method of retaining many partial solutions during a partitioning strategy can be applied to maximum-parsimony methods, and perhaps to maximum-likelihood-based approaches as well. We are currently developing a broader generalization of the approach that can be applied to character-based, rather than distance-based, cost criteria.

Acknowledgments

We wish to thank Dr. Douglas Taylor for his comments on our draft. This research was supported by a grant from the National Library of Medicine (LM04961). G.R. received additional support from an NSF Young Investigator Award and a Packard Foundation Fellowship.

LITERATURE CITED

ALPERT, C., and A. B. KAHNG. 1995. Recent directions in netlist partitioning: a survey. Integration VLSI J. **19**:1–81.

BANDELT, H. J., and A. DRESS. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. **1**:242–252.

DAY, W. 1987. Computational complexity of inferring phylogenies from dissimilarity matrices. Bull. Math. Biol. **49**:461–467.

DAYHOFF, M. O. 1978. Atlas of protein sequence and structure. Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

FELSENSTEIN, J. 1982. Numerical methods for inferring evolutionary trees. Q. Rev. Biol. **57**:379–404.

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

———. 1993. PHYLIP: phylogeny inference package. Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155**:279–284.

FOULDS, L. R., and R. L. GRAHAM. 1982. The steiner problem in phylogeny is NP-complete. Adv. Appl. Math. **3**:43–49.

HUELSENBECK, J. P. 1995. The performance of phylogenetic methods in the four-taxon case. Syst. Biol. **44**:17–48.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

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.

KUMAR, S. 1996. A stepwise algorithm for finding minimum evolutionary trees. Mol. Biol. Evol. **13**:584–583.

LEITNER, T., D. ESCANILLA, C. FRANZEN, M. UHLEN, and J. ALBERT. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. USA **93**:10864–10869.

MADDISON, D. R. 1991. African origin of human mitochondria DNA reexamined. Syst. Zool. **40**:355–363.

PENNY, D., and M. D. HENDY. 1985. The use of tree comparison metrics. Syst. Zool. **34**:75–82.

PENNY, D., M. A. STEEL, P. J. WADDELL, and M. D. HENDY. 1995. Improved analyses of human mtDNA sequences support a recent African origin for Homo sapiens. Mol. Biol. Evol. **12**:863–882.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

RZHETSKY, A., and M. NEL. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9**:945–967.

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.

STEEL, M. A., and M. D. HENDY. 1993. Distributions of tree comparison metrics—some new results. Syst. Biol. **42**:126–141.

STUDIER, J., and K. KEPPLER. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol. **5**: 729–731.

SWOFFORD, D. 1996. PAUP: phylogenetic analysis using parsimony (and other methods). Version 4.0 (test version). Sinauer, Sunderland, Mass.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514. *In* D. M. HILLS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

WILSON, A. C. E., E. A. ZIMMER, E. M. PRAGER, and T. D. KOCHER. 1989. The hierarchy of life, restriction mapping in the molecular systematics of mammals: a retrospective salute. Elsevier Press, Amsterdam.

MICHAEL HENDY, reviewing editor

Accepted February 18, 1999